

Introductory Physics II

Electricity, Magnetism and Optics

by

Robert G. Brown

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

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 cowed, 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	xi
Preface	xi
Textbook Layout and Design	xii
 II: Getting Ready to Learn Physics	 5
Preliminaries	5
See, Do, <i>Teach</i>	5
Other Conditions for Learning	9
Your Brain and Learning	15
How to Do Your Homework Effectively	20
The Method of Three Passes	24
Mathematics	25
Summary	26
Homework for Week 0	26
 III: Electrostatics	 31
Week 1: Discrete Charge and the Electrostatic Field	31
1.1: Charge	34
1.2: Coulomb's Law	38
1.3: Electrostatic Field	39
1.4: Superposition Principle	41
Example 1.4.1: Field of Two Point Charges	44
1.5: Electric Dipoles	47
Homework for Week 1	50
 Week 2: Continuous Charge and Gauss's Law	 63
2.1: The Field of Continuous Charge Distributions	64

Example 2.1.1: Circular Loop of Charge	67
Example 2.1.2: Long Straight Line of Charge	69
Example 2.1.3: Circular Disk of Charge	70
Example 2.1.4: Advanced: Spherical Shell of Charge	72
2.2: Gauss's Law for the Electrostatic Field	76
2.3: Using Gauss's Law to Evaluate the Electric Field	80
Example 2.3.1: Spherical: A spherical shell of charge	81
Example 2.3.2: Electric Field of a Solid Sphere of Charge	82
Example 2.3.3: Cylindrical: A cylindrical shell of charge	85
Example 2.3.4: Planar: A sheet of charge	87
2.4: Gauss's Law and Conductors	88
Properties of Conductors	88
Example 2.4.1: Field and Charge Distribution of a Blob of Conductor	90
Example 2.4.2: Two Thick Plates Plus Wires (Capacitor)	91
Creating Charged Objects	92
Homework for Week 2	95
Week 3: Potential Energy and Potential	105
3.1: Electrostatic Potential Energy	106
3.2: Potential	107
3.3: Superposition	108
Deriving or Computing the Potential	109
3.4: Examples of Computing the Potential	110
Example 3.4.1: Potential of a Dipole on the x -axis	110
Example 3.4.2: Potential of a Dipole at an Arbitrary Point in Space	112
Example 3.4.3: A ring of charge	114
Example 3.4.4: Potential of a Spherical Shell of Charge	116
Example 3.4.5: Advanced: Spherical Shell of Charge	117
Example 3.4.6: Potential of a Uniform Ball of Charge	118
Example 3.4.7: Potential of an Infinite Line of Charge	122
Potential of an Infinite Plane of Charge	123
3.5: Conductors in Electrostatic Equilibrium	123
Charge Sharing	124
3.6: Dielectric Breakdown	125
Homework for Week 3	127
Week 4: Capacitance	131
4.1: Capacitance	133

Example 4.1.1: Parallel Plate Capacitor	134
Example 4.1.2: Cylindrical Capacitor	136
Example 4.1.3: Spherical Capacitor	137
4.2: Energy of a Charged Capacitor	137
Energy Density	138
4.3: Adding Capacitors in Series and Parallel	139
4.4: Dielectrics	142
Example 4.4.1: The Lorentz Model for an Atom	143
Dielectric Response of an Insulator in an Electric Field	145
Dielectrics, Bound Charge, and Capacitance	149
Homework for Week 4	154
Week 5: Resistance	159
5.1: Batteries and Voltage Sources	162
Chemical Batteries	162
The Symbol for a Battery	163
5.1.1: Batteries and Renewable Energy	164
5.2: Resistance and Ohm's Law	165
A Simple Linear Conduction Model	166
Current Density and Charge Conservation	167
Advanced: Differential Form and Maxwell's Equations	169
The Drude Model	170
Ohm's Law	174
5.3: Resistances in Series and Parallel	177
Series	177
Parallel	178
5.4: Kirchhoff's Rules and Multiloop Circuits	179
Kirchhoff's Loop Rule	180
Kirchhoff's Junction Rule	180
Example 5.4.1: The Internal Resistance of a Battery	181
Example 5.4.2: A Multiloop Resistance Problem	183
5.5: RC Circuits	185
Example 5.5.1: Discharging Capacitor	185
Example 5.5.2: Charging Capacitor	187
Homework for Week 5	190

IV: Magnetostatics 197

Week 6: Moving Charges and Magnetic Force 197

6.1: Magnetic Force versus Magnetic Field	198
6.2: Magnetic Force on a Moving Point Charge	198
Example 6.2.1: A Charged Particle Moving in a Uniform Magnetic Field	199
Example 6.2.2: The Cyclotron	200
Example 6.2.3: Cloud Chamber	201
Example 6.2.4: Region of Crossed Fields	202
Example 6.2.5: Thomson's Apparatus for measuring e/m	203
Example 6.2.6: The Mass Spectrometer	207
Example 6.2.7: The Hall Effect	209
6.3: The Magnetic Force on Continuous Currents	210
Example 6.3.1: The Magnetic Force and Torque on a Rectangular Current Loop (Magnetic Dipole)	211
Example 6.3.2: The Magnetic Moment of an <i>Arbitrary</i> Plane Current Loop	213
0.1 Potential Energy of a Magnetic Dipole	214
Example 6.3.3: The Magnetic Moments of Rotating Charged Objects	215
Example 6.3.4: The Precession of Magnetic Moments: Magnetic Resonance	217
6.4: Spin Echoes and Magnetic Resonance Imaging	219
Homework for Week 6	222

Week 7: Sources of the Magnetic Field 227

7.1: Gauss's Law for Magnetism	228
Magnetic Flux	230
7.2: The Magnetic Field of a Point Charge	230
Finite Field Propagation Speed for E and B	231
Violation of Newton's Third Law	232
7.3: The Biot-Savart Law	233
7.4: Examples of Using the Biot-Savart Law to Find the Magnetic Field	235
Example 7.4.1: Magnetic Field of a Straight Wire Segment	235
Example 7.4.2: Field of a Circular Loop on its Axis	237
Example 7.4.3: Field of a Revolving Ring of Charge on its Axis	238
7.5: Ampere's Law	240
7.6: Applications of Ampere's Law	243
Example 7.6.1: Cylindrical Current Density – Infinitely Long Thin Wire	243
Example 7.6.2: Cylindrical Current Density – Field of an Infinitely Long Thick Wire	244
Example 7.6.3: The Solenoid	245
Example 7.6.4: Toroidal Solenoid	247

Example 7.6.5: Infinite Sheet of Current	248
7.7: Summary	249
Homework for Week 7	251
V: Electrodynamics	257
Week 8: Faraday's Law and Induction	257
8.1: Magnetic Forces and Moving Conductors	259
8.2: The Rod on Rails	261
Problem and Solution	263
8.3: Faraday's Law	265
8.4: Lenz's Law	266
0.0.1 Lenz's Law for changing C	267
0.0.2 Lenz's Law for changing B (magnitude)	268
0.0.3 Lenz's Law for changing \vec{B} or \hat{n} direction	268
Example 8.4.1: Wire and Rectangular Loop – Direction Only	269
Example 8.4.2: Rectangular Loop Pulled from Field	271
8.5: More Rod on Rails Problems	271
Example 8.5.1: Rod on Rails with Battery	271
8.6: Inductance	273
Example 8.6.1: The Mutual Inductance of a Wire and Rectangular Current Loop	276
8.7: Self-Induction	278
Example 8.7.1: The Self-Inductance of the Solenoid	278
Example 8.7.2: Toroidal Solenoid	280
Example 8.7.3: Coaxial Cable	281
8.8: LR Circuits	282
Power	284
8.9: Magnetic Energy	285
Example 8.9.1: Energy in a Toroidal Solenoid	286
8.10: Eddy Currents	286
8.11: Magnetic Materials	288
Diamagnetism	288
Superconductors	290
Paramagnetism	290
Ferromagnetism and Antiferromagnetism	291
The Curie Temperature and Neel Temperature	293
Magnetism, Concluded	294

Homework for Week 8	297
Week 9: Alternating Current Circuits	303
9.1: Introduction: Alternating Voltage	312
Electrical Distribution True Facts	313
The Transformer	314
Power Transmission	316
9.2: AC Circuits	318
Non-driven LC circuit	318
Non-driven LRC circuit	319
A Harmonic AC Voltage Across a Resistance R	321
A Harmonic AC Voltage Across a Capacitance C	321
A Harmonic AC Voltage Across an Inductance L	322
The Series LRC Circuit	323
Power in a Series LRC Circuit	325
The Parallel LRC Circuit	328
The AM Radio and Bandwidth	331
Homework for Week 9	336
Week 10: Maxwell's Equations and Light	341
Ampere's Law and the Maxwell Displacement Current	346
Example 10.0.1: The Magnetic Field Inside a Parallel Plate Capacitor	350
10.1: Maxwell's Equations for the Electromagnetic Field: The Wave Equation	352
10.1.1: Accelerating Charge	353
10.1.2: The Wave Equation	354
10.2: Light as a Harmonic Wave	358
10.3: The Poynting Vector	360
10.4: Radiation Pressure and Momentum	362
Homework for Week 10	364
I Optics	369
Week 11: Light	371
11.1: The Speed of Light	374
11.2: The Law of Reflection	374
11.3: Snell's Law	376
Fermat's Principle	377
Total Internal Reflection, Critical Angle	380

Dispersion	381
11.4: Polarization	382
Unpolarized Light	382
Linear Polarization	382
Circularly Polarized Light	383
Elliptically Polarized Light	384
Polarization by Absorption (Malus's Law)	384
Polarization by Scattering	385
Polarization by Reflection	385
Polaroid Sunglasses	386
11.5: Doppler Shift	386
Moving Source	387
Moving Receiver	388
Moving Source and Moving Receiver	388
Homework for Week 11	389
Week 12: Lenses and Mirrors	393
12.1: Vision and Plane Mirrors	395
12.2: Curved Mirrors	397
12.3: Ray Diagrams for Ideal Mirrors	400
12.4: Lenses	403
12.5: The Eye	405
12.6: Optical Instruments	408
The Simple Magnifier	408
Telescope	409
Microscope	411
Homework for Week 12	414
Week 13: Interference and Diffraction	417
13.1: Harmonic Waves and Superposition	422
13.1.1: Hot Sources and Wave Coherence	423
13.1.2: Combining Coherent Harmonic Waves	426
13.2: Interference from Two Narrow Slits	427
13.3: Interference from Three Narrow Slits	430
13.4: Interference from 4, 5, ... N Narrow Slits	433
13.5: The Diffraction Grating – Rayleigh's Criterion for Resolution	436
13.5.1: Rayleigh's Criterion for Resolution	438
13.5.2: Resolving Power	438

13.6: Diffraction	439
13.7: Diffraction Minima, Heuristic Rule	441
13.8: Exact Solution to Diffraction by a Single Slit	442
Example 13.8.1: Diffraction Pattern of a Slit of Width $a = 4\lambda$	447
13.9: Two Slits of Finite Width	447
Example 13.9.1: Two Slits of Separation $d = 8\lambda$ and width $a = 4\lambda$	448
13.10: Diffraction Through Circular Apertures – Limitations on Optical Instruments . . .	449
13.11: Thin Film Interference	452
13.11.1: Phase Shift Due to Path Difference <i>in the Thin Film!</i>	453
13.11.2: Phase Shifts Due to Reflections at the Surfaces	454
13.11.3: No Relative Phase Shift from Surface Reflections	455
13.11.4: A Relative Phase Shift of π from Surface Reflections	455
13.11.5: The Limits of <i>Very</i> Thin Films	456
Homework for Week 13	458

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, a separate supplementary text intended *specifically to help students of introductory physics quickly and efficiently review the required math* is being prepared as a companion volume to all semesters of introductory physics. Indeed, it should really be quite useful for any course being taught with any textbook series and not just this one.

This book is located here:

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

and I *strongly suggest* that all students who are reading these words preparing to begin studying physics pause for a moment, visit this site, and either download the pdf or bookmark the site.

Note that *Week 0: How to Learn Physics* 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).

II: Getting Ready to Learn Physics

Preliminaries

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, detached 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 initial 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 ass¹ 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 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}

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

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 constructed, 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 above².

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 *makes sense* of a sort,

²1357986420 (one, two, three, quit and do something else for one minute...)

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, epiphinic 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.

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

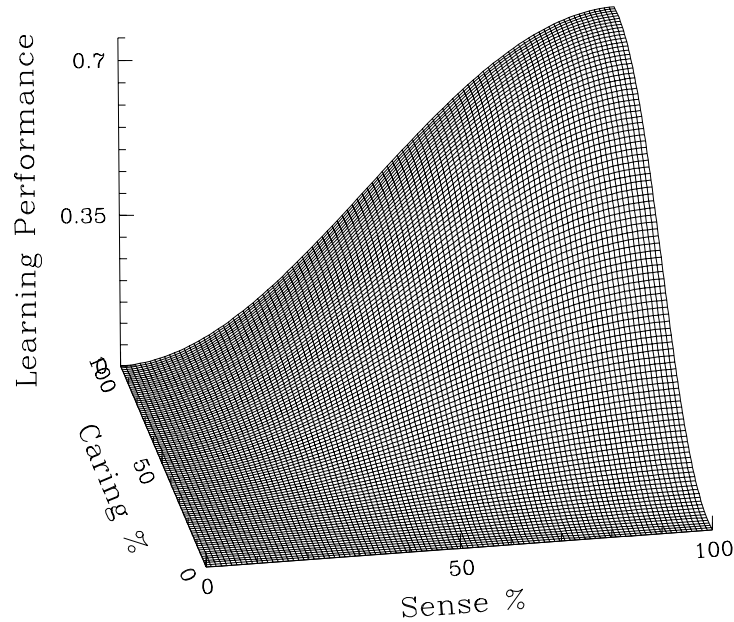


Figure 1: Relation between sense, care and learning

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 index³ tends 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 bed⁴

Learning requires adequate *sleep*. Sure this is tough to manage at college – there are no parents

³Wikipedia: http://www.wikipedia.org/wiki/glycemic_index.

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...

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 ADD⁵ 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 needs⁶ 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, 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

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

⁶Wikipedia: 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.

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 machinery that provided it with sight, sound, and touch in such a way that “you” remain⁷. 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 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 brain⁸ 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.

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

⁸Wikipedia: <http://www.wikipedia.org/wiki/brain>.

⁹Wikipedia: http://www.wikipedia.org/wiki/cerebral_hemisphere.

¹⁰Wikipedia: http://www.wikipedia.org/wiki/Cerebral_cortex.

¹¹Wikipedia: http://www.wikipedia.org/wiki/Neural_network.

- 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!
- 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.
- Intelligence¹⁵, 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

¹²Wikipedia: <http://www.wikipedia.org/wiki/Neurons>.

¹³Wikipedia: <http://www.wikipedia.org/wiki/axon>.

¹⁴Wikipedia: <http://www.wikipedia.org/wiki/neurotransmitters>.

¹⁵Wikipedia: <http://www.wikipedia.org/wiki/intelligence>.

¹⁶Wikipedia: http://www.wikipedia.org/wiki/flynn_effect.

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 memory¹⁷ 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 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 objects¹⁸. 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

¹⁷Wikipedia: <http://www.wikipedia.org/wiki/memory>. Several items in a row are connected to this page.

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

term memory is encoded *completely differently* from short term or sensory/immediate 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 picked²⁰, 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 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

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!

²¹ Wikipedia: <http://www.wikipedia.org/wiki/hippocampus>.

(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 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.

²²Wikipedia: <http://www.wikipedia.org/wiki/amygdala>.

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* desirable 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 crossword 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 *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.

²³Note 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.

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 exercises 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 the actual material for *this* course.

Mathematics

Physics, as was noted in the preface, requires a solid knowledge of all mathematics through calculus. That’s right, the whole nine yards: number theory, algebra, geometry, trigonometry, vectors, differential calculus, integral calculus, even a smattering of differential equations. Somebody may have told you that you can go ahead and take physics having gotten C’s in introductory calculus, perhaps in a remedial course that you took because you had such a hard time with precalc or because you failed straight up calculus when you took it.

They lied.

Sorry to be blunt, but that’s the simple truth. If you are not competent enough in math *right this minute* to know, or be able to find without help the answers *instantly* to the following selection of questions:

- What are the two values of α that solve $\alpha^2 + \frac{R}{L}\alpha + \frac{1}{LC} = 0$?
- What is $Q(r) = \int_0^r \frac{\rho}{R} r'^3 dr'$?
- What is $\frac{d \cos(\omega t + \delta)}{dt}$?
- What are the x and y components of a vector of length A that makes an angle of θ with the positive x axis (proceeding, as usual, counterclockwise for positive θ)?
- What is the cross product of the two vectors $\vec{r} = r_x \hat{x}$ and $\vec{F} = F_y \hat{y}$ (magnitude and direction)?
- What is the inner/dot product of the two vectors $\vec{A} = A_x \hat{x} + A_y \hat{y}$ and $\vec{B} = B_x \hat{x} + B_y \hat{y}$?

then you are going to have to *add* to the burden of learning physics per se the burden of learning, or re-learning, all of the basic mathematics that would have permitted you to answer them easily. These are not idly selected examples, either; they are all things you will have to in fact do early and *often* in this course!

My strong *advice* to you, if you are now feeling the cold icy grip of panic because in fact you are signed up for physics using this book and you couldn't answer *any* of these questions, is to drop the course and go study math until you really master it, *then* come back and take physics. Seriously dude (or dudess). Or resign yourself to a life of misery and enormously hard work just to pass.

Don't go blaming the course, the teacher, this textbook, or anybody but yourself if you proceed unprepared and then fail or suffer – You Have Been Warned.

So, what if you could do *almost* all of these short problems (and can at the very least remember the tools, like the Quadratic Formula, that you were *supposed* to use to solve them if you could remember them? What if you have no choice but to take physics now, and are just going to have to do your best and relearn the math as required along the way? What if you did in fact understand math pretty well once upon a time and are sure it won't be *much* of an obstacle, but you really would like a review, a summary, a listing of the things you need to know someplace handy so you can instantly look them up as you struggle with the problems that uses the math it contains? What if you are (or were) *really good* at math, but want to be able to look at derivations or reread explanations to bring stuff you learned right back to your fingertips once again?

For all of these latter cases, for students of the course in general, I provide the following online (free!) book: ***Mathematics for Introductory Physics***. It is located here:

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

It is a work in progress, and is quite possibly still somewhat incomplete, but it should help you with a lot of what you are missing, and if you let *me* know what you are missing that you didn't find there, I can work to add it!

I would strongly advise all students of introductory physics (any semester) to visit this site *right now* and bookmark it or download the PDF, and to visit the site from time to time to see if I've posted an update. It is on my back burner, so to speak, until I finish the actual physics texts themselves that I'm working on, but I will still add things to them as motivated by my own needs teaching courses using this series of books.

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.

Homework for Week 0

Problem 1.

Skim read this entire section (Week 0: How to Learn Physics), 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. Write a short essay (say, three paragraphs) describing 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.

Problem 2.

Skim-read the entire content of *Mathematics for Introductory Physics* (linked above). Identify things that it covers that you don't remember or don't understand. Pick one and learn it.

Problem 3.

Apply the *Method of Three Passes* to *this* homework assignment. You can either write three short essays or revise your one essay three times, trying to improve it and enhance it each time for the first problem, and review both the original topic and any additional topics you don't remember in the math review problem. On the *last* pass, write a short (two paragraph) essay on whether or not you found multiple passes to be effective in helping you remember the content.

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 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?

III: Electrostatics

Week 1: Discrete Charge and the Electrostatic Field

- **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:

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

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.

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

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{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{r}_0 in the limit that the test charge vanishes:

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

or

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

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

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

- **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:

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

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

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

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:

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

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

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

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

Associated with this torque is the following potential energy:

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

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

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

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*²⁴ 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 Wikipedia: http://www.wikipedia.org/wiki/Triboelectric_effectTriboelectric effect. 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 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*²⁵. Benjamin 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 from one object that otherwise contained equal, balanced amounts of both signs of

²⁴Wikipedia: http://www.wikipedia.org/wiki/electric_charge.

²⁵Wikipedia: 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.

charge, to another, leaving behind a surplus of the other sign. 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. In 1756 Franklin was elected as a Fellow of the Royal Society, which in some ways was the “heart” of the Enlightenment, and remained engaged in natural philosophy (as science was then called) for most of the rest of his life, but his energies from then on were largely diverted to politics.

Many people 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). However, only Coulomb, the inventor of a very sensitive ***torsional balance***, was able to use the balance and his ability to precisely divide charges to precisely demonstrate the correctness of the inverse square law hypothesis and make electrostatics quantitative.

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 use well to know a some of the most important True Facts about charge.

Experimentally, objects can carry a (net) charge q when “electrified” various ways (for example by rubbing materials together). Charge comes in two flavors, $+$ and $-$, but most matter is approximately charge-neutral most of the time. Consequently, as Benjamin Franklin observed, 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).

“Like” charges exert a long range (action at a distance) repulsive force on one another. “Unlike” charges attract. 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 throughout 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). *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*²⁶. Later we will learn to write this conservation law mathematically in terms of the *flux of the current density*, but since we do haven’t 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.

Experimentally, we can readily see that charge can be moved around in very large to extremely small quantities. A natural question is then: Can we continue dividing charge indefinitely, and move an *infinitesimal* 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 it as being a “fluid” that could be moved around in arbitrary amounts.

²⁶Later in the study of physics you may learn of interactions that lead to e.g. *pair production* (or annihilation) – the simultaneous creation (destruction) of a positron-electron pair, for example. 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” in the law. At the “everyday” energies of normal matter at normal temperatures and absent antimatter, one pretty much can ignore this sort of thing and charge is *individually* conserved at the discrete particle level.

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

Table 1: Charge and Mass of First Generation Fermions

However, just as a fluid is itself microscopically particulate, composed of quantized elementary particles, the “elementary” charge (associated with these elementary particles that are the building blocks of all matter) has experimentally turned out to be discrete and essentially indivisible. Indeed, we characterize elementary particles *by* a unique signature consisting of their (rest) mass, their charge, and other measurable properties.

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*²⁷, which carry a charge that is quantized in units of $e/3$, where $e = 1.6 \times 10^{-19} \text{ C}$. The other family are called *leptons*²⁸ which carry a charge that is quantized in units of e itself.

Table 1 summarizes the names and charge properties of the first generation of the quarks and leptons. Note that quarks come in units of $2e/3$ and $-e/3$, 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* – where the neutron is also made up of three quarks: *udd*. We only see particles with a net charge quantized in units of $\pm e$ outside of a nucleon.

Protons are quite massive – they have a rest mass around $938.3 \text{ MeV}/c^2$ ($1.67 \times 10^{-27} \text{ kg}$), almost 2000 times larger than that of an electron at $0.511 \text{ MeV}/c^2$ ($9.11 \times 10^{-31} \text{ kg}$). Neutrons are just a hair more massive than a proton ($939.6 \text{ MeV}/c^2$). 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*. Finally, atoms in turn are “glued” together by electrostatic forces to form molecules, and molecules often stick together to form bulk matter.

As you proceed in your studies in this course, you should keep a *simple* picture of an atom in your mind – a very massive and tiny nucleus surrounded more or less symmetrically surrounded by a much larger “cloud” of light, relatively mobile electrons to the point of electrical neutrality, with clusters of atoms bound together into molecules (the object of the study of *chemistry*). This picture will turn out to be enormously useful to us as we seek to understand electronic properties of matter.

Nearly all matter is made up of atoms and hence nothing but protons, neutrons, and electrons. 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 likely to be surrounded by charge or locked in solids into a rigid structure in such a way that it isn’t terribly mobile, although in fluids ionic charge can move around with either sign. In semiconductors the mobile charge can also be electron “holes” – de facto positive charge carriers consisting of regions of electron deficit that move against an otherwise stationary

²⁷Wikipedia: <http://www.wikipedia.org/wiki/quark>.

²⁸Wikipedia: <http://www.wikipedia.org/wiki/lepton>.

electronic background.

Franklin, unfortunately, thought that the flavor of mobile charge in ordinary conductors was *positive*. In fact, as noted, 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.

Note that all of these elementary charges are quite tiny in terms of their mass and physical extent compared to bulk matter. There is therefore a *lot* of charge in nearly any macroscopic piece of 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* in the nuclei of the atoms that make it up, and note that for every proton there is generally an electron. The inverse of the mass of a proton is thus a good (approximate) measure of the number of charges per unit mass – around 5×10^{26} charges per kilogram of matter! Even a *microgram* (a billionth of a kilogram) of matter thus has well over ten million billion charges.

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 unnecessary – with so many objects, surely an *average* would do for most purposes! We will therefore have frequent cause to “coarse grain” our description of matter – to *ignore* the discrete particulate nature of charge and average out the *total* charge ΔQ in a *finite but very small* volume of matter ΔV . By choosing ΔV small enough that we can treat it like a volume differential but large enough that it contains a lot of charge, we can define a charge *density*. Similarly, we can associate charge densities with two dimensional sheets of matter (for example, a charged piece of paper or metal plate) or one dimensional lines of matter (for example, a wire or piece of fishing line). We summarize this (and define the symbols most often used to represent charge) as:

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

In all of these forms, it is better indeed to think of charge as being the “fluid” that Franklin imagined it to be!

The last property associated with charge that we wish to mention 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 molecules tightly enough that ordinary electric fields cannot remove them. Surplus charge placed on an insulator tends to remain where you put it. Vacuum is an insulator, as is air, although neither is a *perfect* insulator. 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, at least one electron per atom or molecules is only *weakly* bound to its parent and can easily be pushed from one

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* is 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 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 they exist, as semiconductors are *extremely important* to our society. In particular, semiconductors are used in three critical ways: they are used to make diodes (which we will indeed study when we talk of rectification in AM radios), as amplifiers (transistors) (used to make the music adjustably loud enough to listen to), 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/semiconductors> for a more complete discussion.

From this you can see that charge is indeed ubiquitous. We (and everything around us) are made up of charged particles – even the neutral neutrons in the nuclei that make up most of our mass are made up of charged particles. What holds atoms together? What keeps atoms apart? It is time to learn about one of the most important force laws in the Universe, the one that is perhaps most responsible for chemistry and biology.

1.2: Coulomb's Law

Coulomb's Law is very simple. If one charges various objects (for example, two conducting balls suspended from an insulating string so that they are near to one another but not touching) and measures the deflection of the string 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*. We therefore write the force acting *on* charge 1 *due to* charge 2 as:

$$\vec{F}_{12} = k_e q_1 q_2 \frac{(\vec{r}_1 - \vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|^3} \quad (1)$$

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

same sign. A perfect rendition of the verbal statement, but now we can *compute* the force in a specific set of *coordinates*.

The constant:

$$k_e = \frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \quad (2)$$

effectively defines the “size” of the unit of charge in terms of the already known SI units of force and length, and obviously will vary if we change to a different set of units²⁹.

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 describes *action at a distance*. We’d like there to be a *cause* for the observed force that is present *where* the force is exerted, and lacking anything better to do we’ll *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.

1.3: Electrostatic Field

The electrostatic field is the supposed cause of the electrostatic force between two charged objects. Each charged object produces a *field* that emanates from the charge and is the *cause* of the force the other charge experiences at any given point in space. This field is supposed to be present everywhere in space whether or not we measure it.

The fundamental definition of electrostatic field produced by a charge q at position \vec{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{r}_0 in the limit that the test charge vanishes:

$$\vec{E} = \lim_{q_0 \rightarrow 0} \frac{\vec{F}}{q_0} \quad (3)$$

or

$$\vec{E}(\vec{r}_0) = kq \frac{(\vec{r}_0 - \vec{r})}{|\vec{r}_0 - \vec{r}|^3} \quad (4)$$

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

$$\vec{E}(\vec{r}) = \frac{kq}{r^2} \hat{r} \quad (5)$$

In general, we’ll work the other way around. First we’ll be given a distribution of charges, from which we must determine the field. With the field known, we can then evaluate the force these charges will exert on *another* (e.g. test) charge placed on the field by means of the following rule:

$$\vec{F} = q\vec{E} \quad (6)$$

A common question that students often ask is: “Why all of the hassle with letting test charges go to zero if you’re just going to divide it out anyway?” 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. If that

²⁹Actually the size of the Coulomb was originally defined in terms of the *Ampere* – the unit of electrical current – and *magnetic* forces. We’ll learn about this in a few weeks when we study magnetism.

charge is not nailed down and can move *at all* in response to the test charge, it would rearrange and thereby *change* the field one is trying to measure. By letting it go to zero, one also causes any disturbance caused by the measurement to go to zero, leaving you with the field that is there in the absence of all charges.

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 fields of many charges simply add right up! 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 it the *Superposition Principle*.

1.4: Superposition Principle

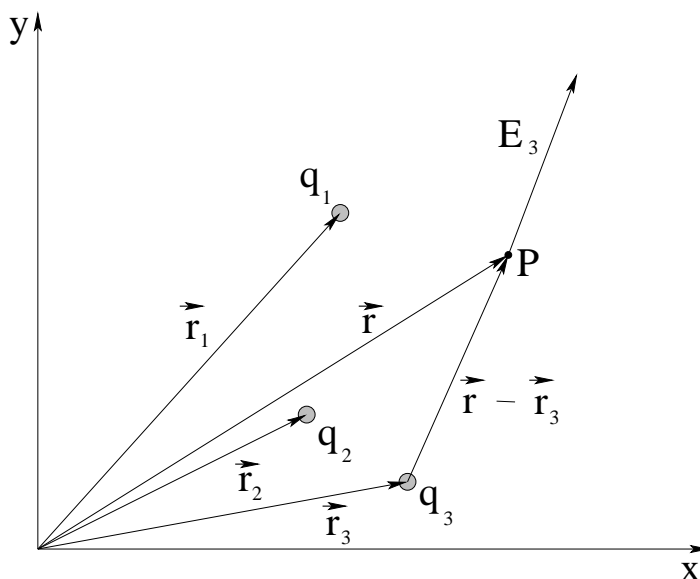


Figure 2: Geometry needed to evaluate the field of many charges. Only the field of the third charge \vec{E}_3 is shown explicitly. Note well the magnitude and direction of the vector $\vec{r} - \vec{r}_3$ – head at \vec{r} , tail at \vec{r}_3 . This is a vector *from* the position of the charge q_3 *to* 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:

$$\vec{E}(\vec{r}) = \sum_i \frac{kq_i(\vec{r} - \vec{r}_i)}{|\vec{r} - \vec{r}_i|^3} \quad (7)$$

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 are *linear* differential equations.

Note that it doesn't have to be that way. There is nothing inherently contradictory about two charges producing a field at a point in space that is *less* than their sum or *more* than their sum. There are examples in physics of interactions that do just that (although this sort of complication, like the “three body forces” that are also excluded by linearity, makes the theories *much* more difficult to solve).

In pure classical physics the field is strictly linear, but in quantum theory the electromagnetic field becomes (in a sense) *nonlinear* at very short distances from elementary charges due to vacuum polarization and in just the right way to “soften” the singularity in certain interactions and be unified with other forces of nature in a single field “theory of everything”. In *this* course, however, we will never ever explore the quantum distance or interaction scales where this sort of thing is an issue, so for us superposition will be a fundamental principle.

As noted above, charge, while discrete, comes in very tiny packages of magnitude e such that matter contains order of 10^{27} charges per kilogram, with roughly equal amounts of positive and negative charge so that most matter is approximately electrically neutral most of the time. When we consider macroscopic objects – ones composed of these enormous numbers of atoms and charges – it therefore makes sense to treat the distribution and motion of charge as if it is *continuously* distributed.

In order to find the electrostatic field produced by a charge density distribution, we use the

superposition principle in *integral* form. Note that the result of this sort of computation will *fail* if we examine \vec{r} inside the material itself very close to one of the constituent discrete charges (where the $1/r^2$ nature of the force guarantees that if you are close enough to a charge, its field will overwhelm the field of all more distant charges) but in general the resulting numbers are both useful and remarkably accurate, accurate as an “average value” even within a material.

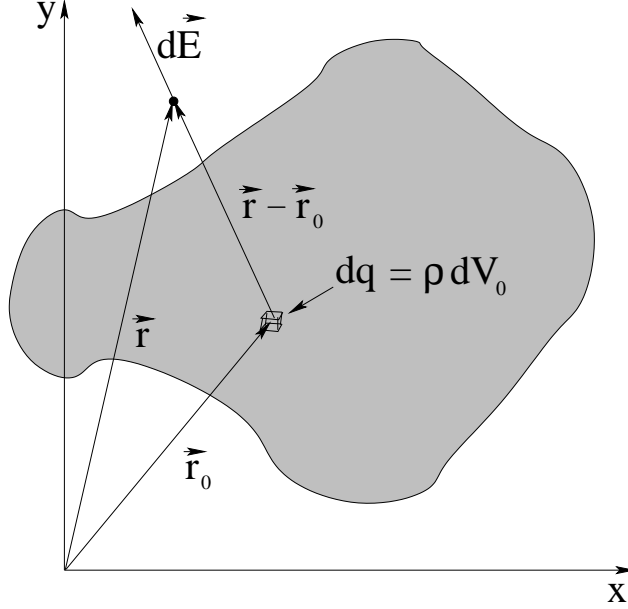


Figure 3: The geometry needed to evaluate the field of a general continuous charge distribution. Note well the similarity to the geometry for a collection of charges, except that the many “point charges” are all chunks of differential volume with charge dq and the “sum” is now an integral.

To write down the integral (and help us remember it) we begin by using the basic equation obtained above for the field of a point charge and apply it to a tiny “chunk” of the charge distribution dq – one small enough to be considered a point-like charge. We write this as the *differential* contribution of the charge to the overall field as follows:

$$d\vec{E}(\vec{r}) = \frac{k_e dq (\vec{r} - \vec{r}_0)}{|\vec{r} - \vec{r}_0|^3} \quad (8)$$

We then use one of the definitions of charge density to convert dq into e.g. $dq = \rho dV_0 = \rho(\vec{r}_0) d^3r_0$:

$$d\vec{E}(\vec{r}) = \frac{k_e \rho(\vec{r}_0) (\vec{r} - \vec{r}_0) d^3r_0}{|\vec{r} - \vec{r}_0|^3} \quad (9)$$

Finally, we integrate both sides of this equation over the entire volume V where $\rho(\vec{r}_0)$ is supported. The resulting integral form is:

$$\vec{E}(\vec{r}) = k_e \int_V \frac{\rho(\vec{r}_0) (\vec{r} - \vec{r}_0) dV_0}{|\vec{r} - \vec{r}_0|^3} \quad (10)$$

for a 3-dimensional (volume) charge distribution,

$$\vec{E}(\vec{r}) = k_e \int_S \frac{\sigma(\vec{r}_0) (\vec{r} - \vec{r}_0) dS_0}{|\vec{r} - \vec{r}_0|^3} \quad (11)$$

for a surface charge distribution on a surface S , and

$$\vec{E}(\vec{r}) = k_e \int_L \frac{\lambda(\vec{r}_0) (\vec{r} - \vec{r}_0) dL_0}{|\vec{r} - \vec{r}_0|^3} \quad (12)$$

for a linear charge distribution on a particular line L .

Because one has to integrate the vector components independently, and since their contribution and geometry can vary as one moves \vec{r} about in space, this integral is remarkably difficult to integrate in the general case for most charge density distributions. We will manage to find a few examples (below) where the difficulty of the integration process is reduced due to the *symmetry* of the charge distribution, which may allow us to cancel (and hence avoid having to do) particular parts of the integrals from symmetry alone, but the methodology overall will be very cumbersome and is rarely used in real physics problems.

Instead in a few chapters we'll derive a similar form, but far more tractable integral form for the electrostatic *potential*, a scalar quantity, and obtain the field (if it is desired at all) by taking the negative gradient of the potential, since vector calculus differentiation is often easier algebraically than vector calculus integration. Even here, however, from a purely practical point of view only very simple and symmetric charge distributions can be solved algebraically, and for most "real world" problems one must resort to using a computer to *numerically integrate* the expressions above, by (for example) computing a direct sum of the fields or potentials in the \sum_i form where each $q_i = \rho\Delta V_i$ for some suitable partitioning of the distribution into a finite number of indexed chunks of size ΔV_i .

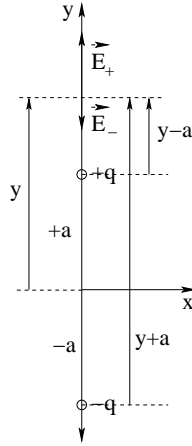
Example 1.4.1: Field of Two Point Charges

Figure 4: Two charges $\pm q$ on the y -axis produce a field that is easy to evaluate at points on the x and y -axis (and not terribly difficult to *approximately* evaluate at all points in space that are “far” from the origin relative to a). This arrangement of charges is called an *electric dipole* and is a very important concept that we will work with extensively below.

Suppose two point charges of magnitude $-q$ and $+q$ are located on the y -axis at $y = -a$ and $y = +a$, respectively. Find the electric field at an arbitrary point on the x and y axis.

The y -axis is quite simple. The field due to the positive charge points directly away from it, hence in the positive y direction at a point $y > a$ and is equal to:

$$\vec{E}_+(0, y) = \frac{k_e q}{|y - a|^2} \hat{y} \quad (13)$$

The field of the negative charge points towards it and is equal to:

$$\vec{E}_-(0, y) = -\frac{k_e q}{|y + a|^2} \hat{y} \quad (14)$$

Hence the total field on the y axis is just:

$$\vec{E}_{\text{tot}}(0, y) = k_e q \left(\frac{1}{|y - a|^2} - \frac{1}{|y + a|^2} \right) \hat{y} \quad (15)$$

The field on the x -axis is a *tiny* bit more difficult. Here the field produced by *each* charge has *both* components. To find the vector field, we must first find the magnitude of the field, then use the *geometry of the picture* to find its x and y components.

Note that the distance from the 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:

$$|\vec{E}(x, 0)| = \frac{k_e q}{r^2} = \frac{k_e q}{(x^2 + a^2)} \quad (16)$$

Look at the right triangle formed by x , a and r . By definition:

$$\cos(\theta) = \frac{x}{r} = \frac{x}{(x^2 + a^2)^{1/2}} \quad (17)$$

$$\sin(\theta) = \frac{a}{r} = \frac{a}{(x^2 + a^2)^{1/2}} \quad (18)$$

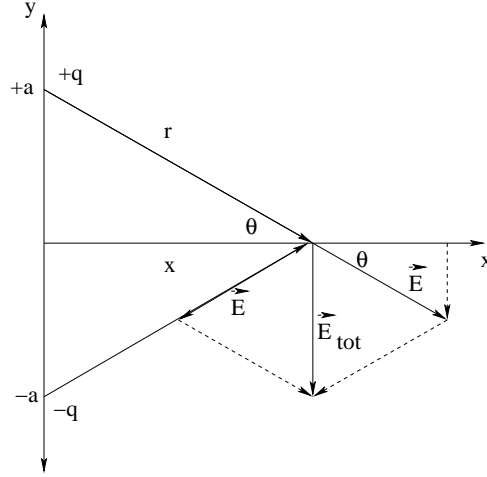


Figure 5: Two charges $\pm q$ on the y -axis produce a field that is (still) pretty easy to evaluate at a points on the x -axis.

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

$$\begin{aligned} E_x &= |\vec{E}| \cos(\theta) = \frac{k_e q}{(x^2 + a^2)} \cdot \frac{x}{(x^2 + a^2)^{1/2}} \\ &= \frac{k_e q x}{(x^2 + a^2)^{3/2}} \end{aligned} \quad (19)$$

and

$$\begin{aligned} E_y &= -|\vec{E}| \sin(\theta) = -\frac{k_e q}{(x^2 + a^2)} \cdot \frac{a}{(x^2 + a^2)^{1/2}} \\ &= -\frac{k_e q a}{(x^2 + a^2)^{3/2}} \end{aligned} \quad (20)$$

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

$$\vec{E}_{\text{tot}}(x, 0) = -2 \frac{k_e q a}{(x^2 + a^2)^{3/2}} \hat{y} \quad (21)$$

In terms of the *electric dipole moment* for this arrangement of charges:

$$\vec{p} = 2qa\hat{y} \quad (22)$$

the field can be expressed as:

$$\vec{E}_{\text{tot}}(x, 0) = -\frac{k_e |\vec{p}|}{(x^2 + a^2)^{3/2}} \hat{y} \quad (23)$$

It is worthwhile to look at the general shape of the dipole field. It is already familiar to any student who has done the simple experiment of sprinkling iron filings onto a sheet of paper sitting on a small bar magnet – the resemblance is not a coincidence, as we shall see.

The electric field and electric potential of a dipole will be of great interest to us over the course the next few weeks. In many cases, the physical dimensions of the dipole ($2a$ in this case) will be *small* compared to x , the distance of the point of observation to the dipole. In this limit, the field or potential produced is that of an *ideal* dipole, or a *point* dipole. We can find the field in the limit that $x \gg a$ very easily by factoring out the larger of the two quantities from the denominator, expressing

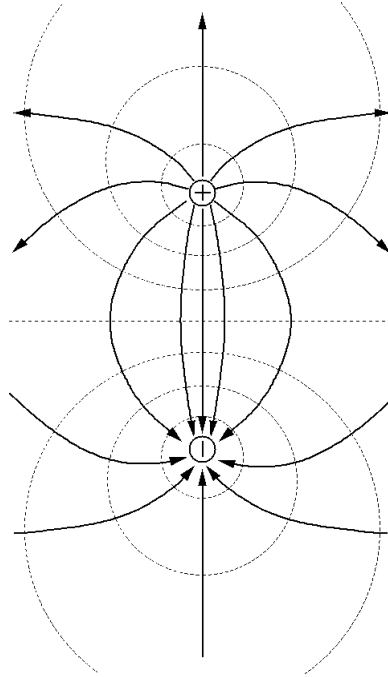


Figure 6: The electric field of a classic electric dipole in the vicinity of the charges. Bear in mind that this figure is a plane cross-section of a three-dimensional, cylindrically symmetric field! The dashed lines are the projections into the plane of the *equipotential surfaces* of this arrangement of charges. As we shall see later, finding the (scalar!) *potential* of an electric dipole is very easy, where finding the field inevitably involves a certain amount of vector annoyance.

the denominator on top (with a negative exponent) in the numerator, and then performing a *binomial expansion* and keeping terms to any desired degree of precision. In this case the process yields:

$$\begin{aligned}
 \vec{E}_{\text{tot}}(x, 0) &= -\frac{k_e |\vec{p}|}{(x^2 + a^2)^{3/2}} \hat{y} \\
 &= -\frac{k_e |\vec{p}|}{x^3} \left(1 + \left(\frac{a}{x}\right)^2\right)^{-3/2} \hat{y} \\
 &\approx -\frac{k_e |\vec{p}|}{x^3} \left(1 - \frac{3}{2} \left(\frac{a}{x}\right)^2 + \dots\right) \hat{y} \\
 &\approx -\frac{k_e |\vec{p}|}{x^3} \hat{y} + \mathcal{O}\left(\frac{1}{x^5}\right)
 \end{aligned} \tag{24}$$

(where the last term is read “plus neglected terms of order $1/x^5$ ”).

As we will see later the field of a point dipole 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 back to the section on math and review both the binomial and Taylor series expansions, as they will be very useful to us as we solve problems and work examples. 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.

1.5: Electric Dipoles

As we just noted, the arrangement of two equal but opposite charges above is called an *electric dipole*³⁰, and dipole fields play an enormously important role in physics. That is because dipolar arrangements of charge are *common* in nature. Let's see why.

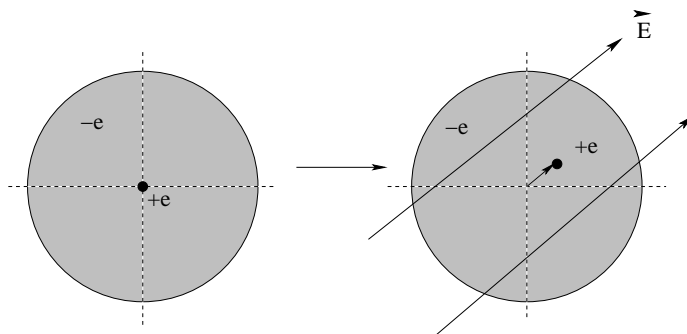


Figure 7: An atom in an electric field *polarizes* as its nucleus is displaced relative to its electron cloud. We will work with a very simple *model* for an atom called the *Lorentz Oscillator Model* that idealizes an atom as a uniform ball of negative charge symmetrically surrounding a small massive positively charged nucleus (such that the total charge is zero). This model works well all the way up to *graduate* electrodynamics to help students understand the general principles of dielectric polarization!

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 produces (as we shall see) no electric field outside the atom. If such an atom is placed in an electric field, the nucleus is pulled one way and the electron cloud is pushed the other way, and while the atom remains electrically neutral the vector fields produced by the positive and negative charges are symmetric about different centers and *no longer precisely cancel*.

In a few weeks we will consider the field produced by the polarized atoms *on average* inside a solid as this field *modifies* the field that polarizes the atoms and we will learn some wonderful things, such as the fact that the natural motion of the charge distribution in this idealized model is to *harmonically oscillate* (hence its name: the *Lorentz Oscillator Model*).

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 – some stuff that seems merely interesting and other stuff that seems as though it might be very important indeed in our efforts to build a rational worldview that explains the world of our everyday experience in simple, intuitive terms.

We will therefore start by modelling the resulting charge distribution of a polarized atom (or any other dipolar system) as a basic 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:

When two electric charges of equal magnitude and opposite sign are bound together, they form an *electric dipole*. To understand the properties of dipoles as “objects”, we will initially presume them to be bound together with a “rigid rod” of some sort so the dipole moment itself *doesn't*

³⁰Wikipedia: <http://www.wikipedia.org/wiki/dipole>.

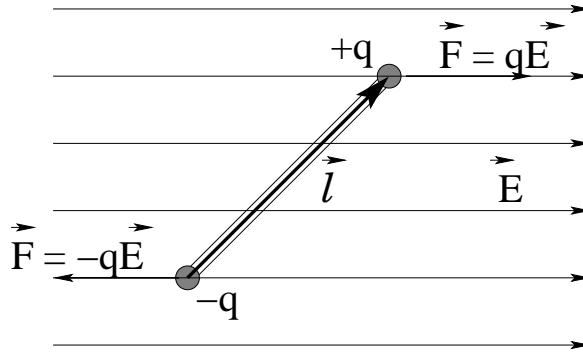


Figure 8: The basic dipole consists of two equal and opposite charges $\pm q$ separated by a vector displacement \vec{l} , in which case the *dipole moment* of the arrangement is defined to be $\vec{p} = q\vec{l}$. Note that we are not in this figure assuming that the E -field is *creating* the dipole; rather we are assuming that it is fixed, with the charges rigidly separated by e.g. a massless rod in between.

change in response to any field one might put them in, although this is clearly only a model and not the reality for most real dipoles bound together by a *non-rigid* force. 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:

$$\vec{p} = q\vec{l} \quad (25)$$

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

In the example above and the homework, we algebraically evaluate the field produced by a dipole along lines of symmetry where the field has a simple form, and qualitatively draw out the general form of the field at *arbitrary* points in space as illustrated in figure 6. The electric field of a “point like” dipole has an extremely characteristic shape and a precisely defined functional form in terms of \vec{p} , although we will find it far simpler to evaluate the electrostatic *potential* of a dipole at an arbitrary point when we get to the appropriate chapter.

At this point, let us consider the *force* and the *torque* exerted by an electric field on a dipole. If an electric dipole is placed in a *uniform* electrical field, the forces on the two poles are equal in magnitude and opposite in direction. The net force on the dipole is therefore zero. Algebraically:

$$\begin{aligned} \vec{F} &= -q\vec{E} + q\vec{E} \\ &= 0 \end{aligned} \quad (26)$$

If the dipole is not aligned or antialigned with the uniform field, however, the field clearly exerts a *torque* on the dipole. The forces form a “couple” (two opposite forces that do not act along the same line), and therefore this torque is independent of our choice of pivot (see *Introductory Physics I* if necessary to review this and other aspects of torque).

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

$$\begin{aligned} \vec{\tau} &= \vec{r} \times \vec{F} \\ &= \vec{l} \times q\vec{E} \\ &= q\vec{l} \times \vec{E} \\ &= \vec{p} \times \vec{E} \end{aligned} \quad (27)$$

(noting that charge is a scalar quantity). This is a very important result; learn this picture and mini-derivation well so you can easily remember and apply it. Since this is the first time this semester

that you have seen a *cross product*, if you have started to forget it needless to say it is a very good idea to backtrack to the math section of this textbook and review its pictorial representation, its algebra and geometry, and of course the good old right hand rule!

Associated with this torque is the following potential energy which is clearly minimized when the dipole moment aligns with the applied field. We look at the picture above, and consider the amount of work done by only the component of the force perpendicular to the arc of motion as we twist the dipole from a position at right angles to the field (where we define the potential energy to be zero) to an arbitrary angle. A bit of consideration and a good picture (see homework) should convince you that:

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

(**Note well!** The force/torque has the *opposite* sign to the angle θ !) or

$$U = -\vec{p} \cdot \vec{E} \quad (28)$$

Note that $U(\theta)$ is minimum (negative) when the dipole is aligned with the field, maximum (positive) when antialigned.

This expression is only generally exact if \vec{p} is a “point dipole”, since it assumes that \vec{E} is at least approximately the same at the two ends of the dipole so the forces form a couple and the energy is strictly due to the torque. More practically, however, it is *usable* (and quite accurate) whenever the dipole is short relative to the scale over which \vec{E} varies, so that the value of \vec{E} “at the position of the dipole” is a well-defined quantity. From this and our general knowledge of intro-level mechanics, we can see that the force on the dipole in a more general *non-uniform* field *should* be:

$$\vec{F} = -\vec{\nabla}U = \vec{\nabla}(\vec{p} \cdot \vec{E}) \quad (29)$$

which can be difficult to compute but is easy to understand³¹.

In our simple model for the dipole above, if the field is *not* uniform then it will in general *not* be equal at the locations of the two charges. In fact, if we let \vec{E} be the field at (say) the location of the negative charge and $\vec{E}' = \vec{E} + \Delta\vec{E}$ at the location of the positive charge, we have:

$$\begin{aligned} \vec{F} &= -q\vec{E} + q\vec{E}' \\ &= -q\vec{E} + q\vec{E} + q\Delta\vec{E} \\ &= q\Delta\vec{E} \\ &= \vec{\nabla}(\vec{p} \cdot \vec{E}) \end{aligned} \quad (30)$$

where the last step, in very rough terms, results from letting $\vec{p} = q\Delta\vec{l}$ (a very short point-like dipole) then $\Delta\vec{E} \approx \Delta\vec{l} \cdot \vec{\nabla}\vec{E}$ is basically the first term of a Taylor series expansion of \vec{E} , where the gradient has to be applied to each component of the field separately. This will be explored further in homework problems.

³¹Students who have never seen the gradient operator $\vec{\nabla}$ before and who are not potential physics or math majors will not be *tested* on this, but are still advised to read and *study* it and to try to understand it, because it actually explains a lot of things very compactly that otherwise (as we have seen and will see further below) are actually *more* difficult to derive and evaluate than the gradient.

Homework for Week 1

Note well that there are “no numbers” in the following problems. Most problems are for “all students of physics”. Some problems are marked with a * as “advanced” and are intended to be assigned primarily to physics majors or engineering students, who are expected to know and use a bit more calculus than life science students, but note well that there is *plenty of calculus in the general problems!* It is impossible to learn and understand physics without calculus; Newton invented calculus *just so he could formulate physics* and this course *teaches* the correct use of algebra, geometry, trigonometry, calculus in general including simple differential equations (e.g. the harmonic oscillator, the wave equation) in the solving of problems.

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:

$$\vec{F}_{ij} = \frac{k_e q_i q_j (\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|^3}$$

(with $k_e = 9 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2$). Needed in problem(s) 4, 5, 6, 7, 8, 9, 10, 11. A core concept!

- Electric Field:

$$\vec{E} = \lim_{q_0 \rightarrow 0} \frac{\vec{F}_0}{q_0} = \frac{k_e q (\vec{r}_0 - \vec{r})}{|\vec{r}_0 - \vec{r}|^3}$$

or of a point charge, located at the origin:

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

Needed in nearly all of the problems.

- This definition ensures that we can find the force on a charge as follows:

$$\vec{F} = q\vec{E}$$

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:

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

or, for a continuous distribution of charge:

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

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{l} is:

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

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:

$$\vec{F} = 0$$

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

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

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

- Yes, we use Newton's Second Law:

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

(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:

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

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

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

(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:

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

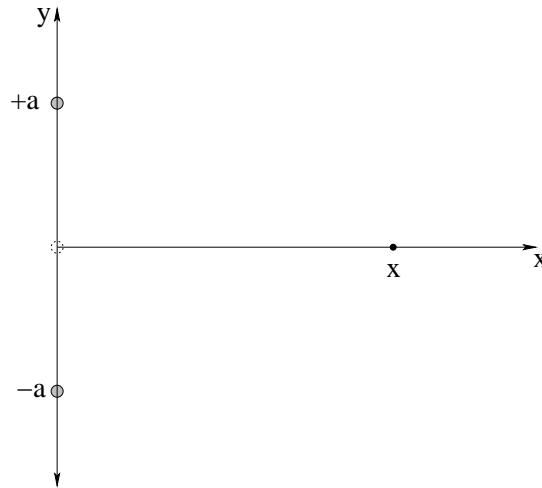
which converges for small Δa (used in problems 3, 5, 6, 11) and the Taylor series of a particular functional form, the **Binomial Expansion**:

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

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

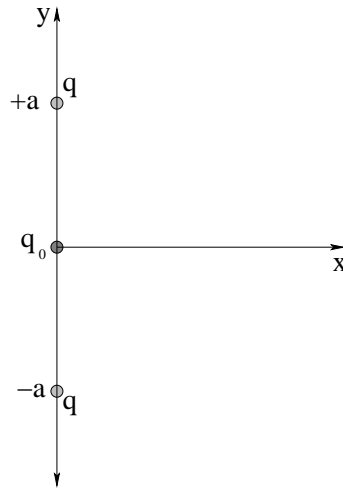
Note well the similarity between this concepts summary *needed for the homework* and the concepts summary that started the chapter. This is no accident; the chapter summary is there at the start for a reason! However, there may be additions or deletions – don't just copy the summary, and **be sure to 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 equal positive charges $+q$ sit at $y = -a$ and $y = +a$.

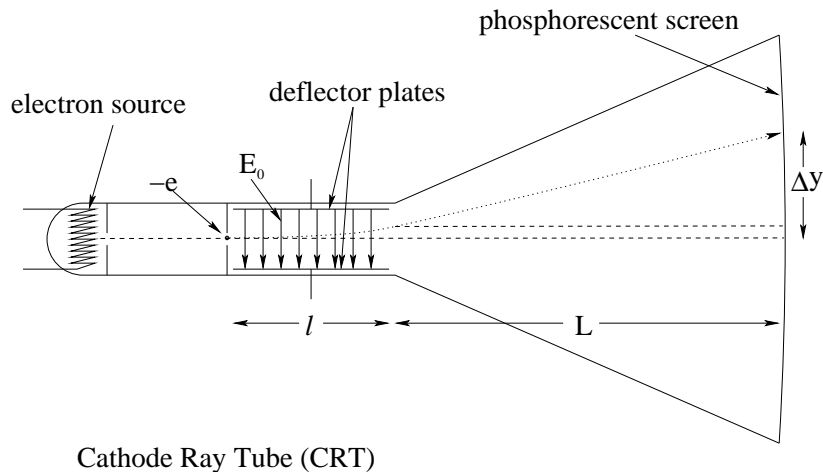
- a) Find the electric field at an arbitrary point on the x axis, and find its asymptotic form when $x \ll a$ (near the origin) and $x \gg a$ (far from the pair of charges). Explain the latter result intuitively.
- b) Repeat for a positive charge $+q$ at $y = +a$ and a negative charge $-q$ at $y = -a$.
- c) Repeat for two equal positive charges $+q$ sitting 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. What might we call the first surviving term in the distant field? (Note that there are *four* monopoles in this distribution.)

Problem 3.

Two equal positive charges are on the y axis, one at $y = +a$ and the other at $y = -a$. The electric field at the origin is zero. A test charge q_0 placed at the origin will therefore be in equilibrium.

- Discuss the stability of the equilibrium for a positive test charge by considering small displacements from equilibrium along the x axis and small displacements along the y axis.
- Repeat part (a) for a negative test charge.
- Find the magnitude and sign of a charge q_0 that when placed at the origin results in a net force of zero on each of the three charges. What will happen if any of the charges are displaced slightly from equilibrium in different directions (is the equilibrium stable, unstable, metastable)?

The point at the origin is called a *saddle point* because the potential there is shaped like a *saddle*, with a smooth minimum along one axis and a smooth maximum along the axis perpendicular to it. Bear this in mind for a couple of weeks until we define and evaluate electrostatic potential!

Problem 4.

Cathode Ray Tube (CRT)

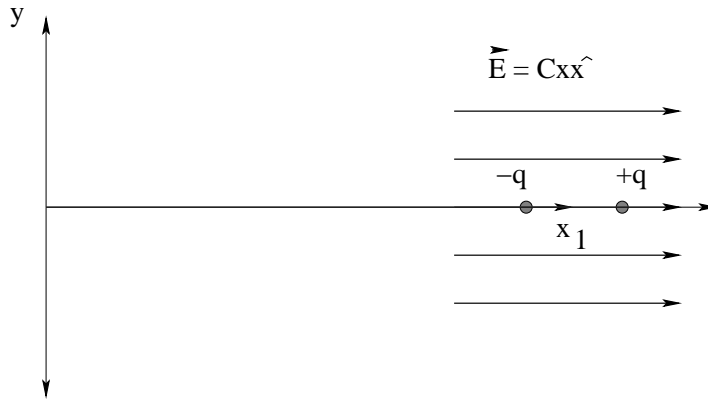
An electron moves to the right with speed v along the axis of a cathode ray tube. There is an electric field $\vec{E} = E_0 \hat{j}$ in the region between the deflection plates, which are of length l , and everywhere else $\vec{E} = 0$. The flat screen is a distance L from the end of the plates. Assume that the electron is moving fast enough that it will not “fall” or hit the deflection plates while crossing the deflection zone (ignore effect of the gravitational force on the electron as it is negligible across the entire distance).

Find Δy , the deflection from the center point where the electron hits the screen. You might want to break the problem up into two parts as the figure hints.

Problem 5.

A ball of known charge q and unknown mass m , initially at rest, falls freely from a height h in a uniform electric field \vec{E} that is directed vertically downward. The ball hits the ground at a speed $v = 2\sqrt{gh}$.

Find m in terms of E , q and g .

Problem 6.

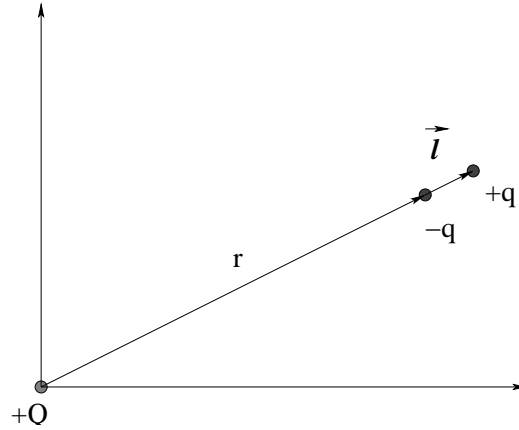
An electric dipole consists of two charges $+q$ and $-q$ separated by a very small distance $2a$. Its center is on the x axis at $x = x_1$, and it points along the x axis in the positive x direction. The dipole is in a nonuniform electric field which is also in the x direction, given by $\vec{E} = Cx\hat{x}$, where C is a constant.

- Find (write down, it's trivial) \vec{p} , the (vector) dipole moment of this electric dipole. Note its magnitude p_x .
- Find the force on the positive charge and that on the negative charge, and show that the net force on the dipole is $C p_x \hat{x}$.
- Show that *in general*, if a dipole lies along the x axis in an electric field in the x direction so that $\vec{p} = p_x \hat{x}$, the net force on the dipole is given approximately by:

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

where the derivative of the field is evaluated at the position of the dipole.

You will probably need to use a Binomial/Taylor expansion to deal with the “ $r \gg L$ ” condition. Your instructor or TA will help you with this if you have no idea how to proceed.

Problem 7.

A positive point charge $+Q$ is at the origin, and a dipole of moment \vec{p} is at a distance r away and pointing in the radial direction (where $r \gg L$, the physical length of the dipole) as shown.

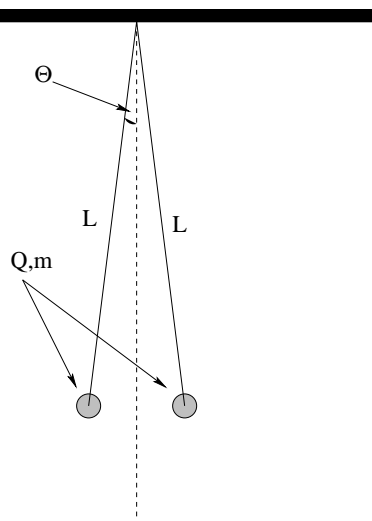
- a) Show that the force exerted on the dipole by the point charge is attractive and has a magnitude

$$F_r \approx \frac{2kQp}{r^3}.$$

- b) Now assume that the dipole is centered at the origin and that a point charge Q is a distance r along the line of the dipole. Using Newton's third law and your result for part (a), show that at the location of the positive point charge the electric field due to the dipole is toward the dipole and has a magnitude of

$$E_r \approx \frac{2kp}{r^3}$$

Again, you will probably need to use a Binomial/Taylor expansion to deal with the " $r \gg L$ " condition. Your instructor or TA will help you with this if you have no idea how to proceed. Or, you might be able to do it by considering this one a special case of the previous problem, if you can mentally rotate coordinate systems...

Problem 8.

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.

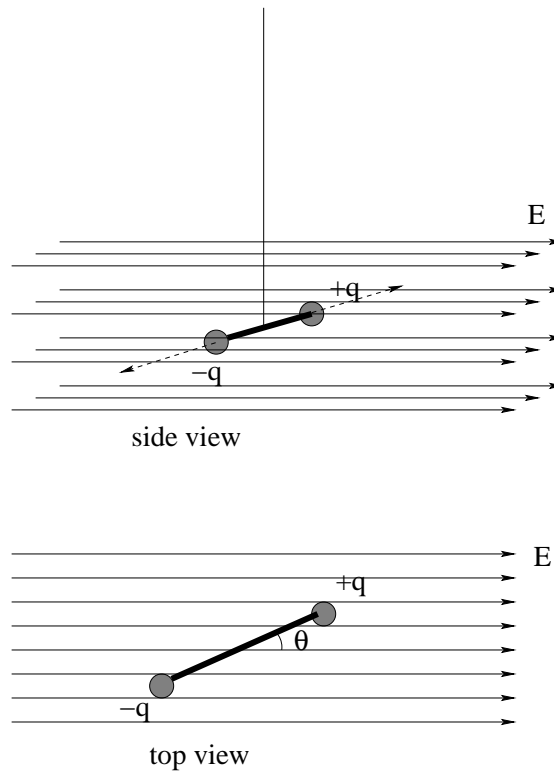
- a) Show that the charge q is given by:

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

where k_e is the electrostatic constant.

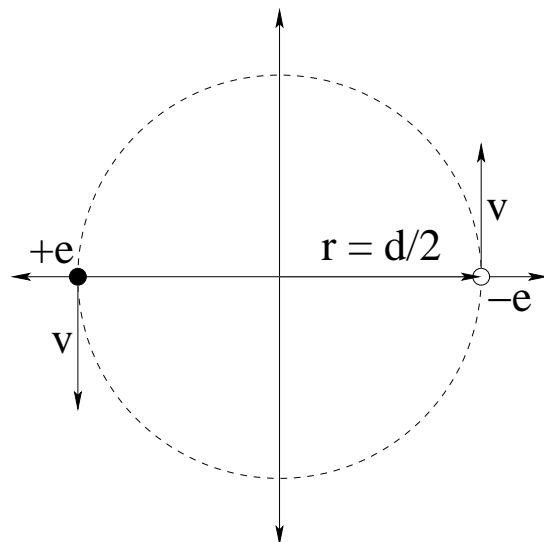
- b) Find q if $m = 10$ grams, $L = 50$ cm, and $\theta = 10^\circ$. You may (as usual) use $g = 10$ m/sec².
 c) What would happen if both charges q equalled 1 Coulomb instead of the tiny charge you obtained in your answer to b)?

Note that numbers are given in this problem primarily to *just once* force you to confront what a reasonable “size” is for macroscopic electric charges in the laboratory. Note well that it is much, much smaller than a Coulomb!

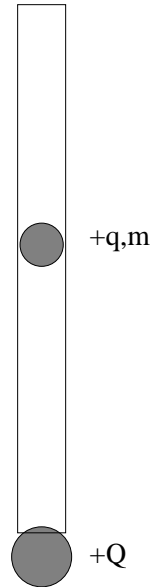
Problem 9.

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

Show that for small values 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. (If you look back at the concepts section, I remind you of the *form* of the simple harmonic oscillator equation – the idea is to transform the equation of motion into this form, at which point you *know the solution and all about the associated motion* from having solved it repeatedly in the first part of the course.)

Problem 10.

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 annihilate one another). Find the speed of each particle v in terms of e , m , k and their separation d . Note well that the circle of their motion has a radius $r = d/2$!

Advanced Problem 11.

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 .

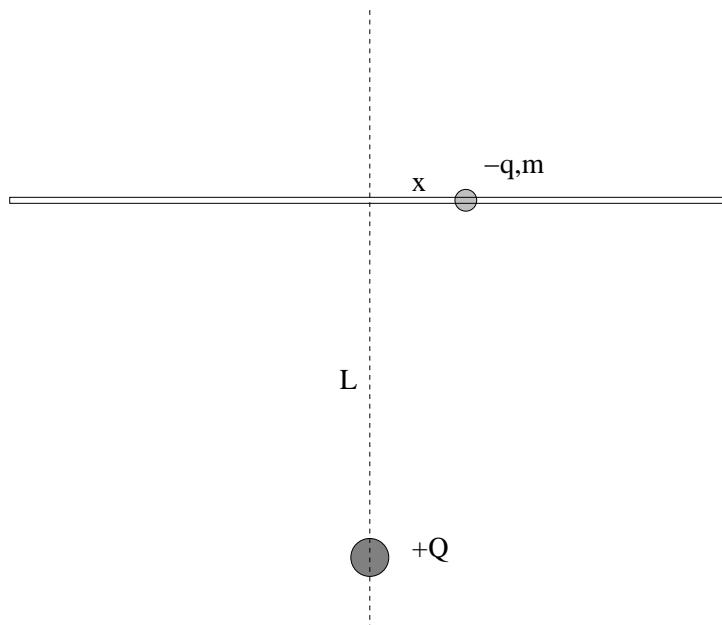
- a) Show that the mass m will be in equilibrium at a height:

$$y_0 = \sqrt{\frac{kqQ}{mg}}$$

- 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:

$$\omega = (2g/y_0)^{1/2}$$

You will need to use expansions to solve this problem.

Advanced Problem 12.

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 .

You will need to use expansions 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:

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

where $dV_0 = d^3r_0$ 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:

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

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

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:

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

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.

- The electric field vanishes inside a conductor *in electrostatic equilibrium* (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).
- All non-neutral charge distributed on a conductor *in electrostatic equilibrium* must reside on the surface.
- The electric field at the *surface* of a conductor *in electrostatic equilibrium* must begin or terminate on the conductor *perpendicular* to the surface. There can be no field component parallel to the surface of a conductor.
- Since the field at the surface of a conductor is \vec{E}_\perp only 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:

$$\vec{E}_\perp = 4\pi k_e \sigma = \frac{\sigma}{\epsilon_0}$$

The field at the surface is directly proportional to the surface charge density!

2.1: The Field of Continuous Charge Distributions

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:

$$\rho = \frac{\Delta Q}{\Delta V} \tag{31}$$

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 with a calculator. Generally, however, 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:

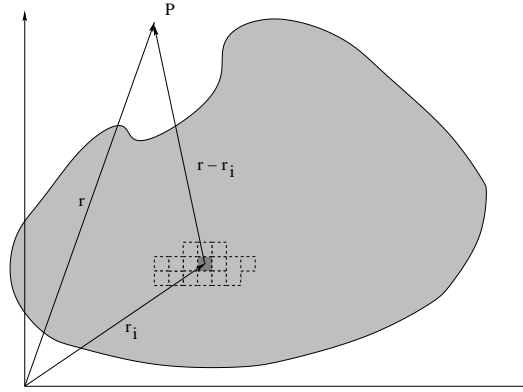


Figure 9: Coarse grained average leading to an integral.

In the amoebic blob shaped object above, we've chopped the whole volume up into little chunks ΔV in size (highly exaggerated in the picture so you can see them). We've tallied up the charge in each block ΔQ , and labelled (in our minds) each block with an index i at position \vec{r}_i . We can then compute the field using the superposition principle at the point P (position \vec{r}) as:

$$\vec{E}_{\text{tot}}(\vec{r}) = \sum_i \frac{k\Delta Q_i}{|\vec{r} - \vec{r}_i|^2} \widehat{(\vec{r} - \vec{r}_i)} \quad (32)$$

As noted, there are too many chunks in the blob for us to sum over. So we pretend that the charge is *continuously distributed* according to:

$$\rho = \lim_{\Delta V \rightarrow 0} \frac{\Delta Q}{\Delta V} = \frac{dQ}{dV} \quad (33)$$

and turn the summation into an *integral* (remember both σ and \int stand for S(um), they are both summation symbols, the latter the one we use for continuous things):

$$\vec{E}_{\text{tot}}(\vec{r}) = \sum_i \frac{k\Delta Q_i}{|\vec{r} - \vec{r}_i|^2} \widehat{(\vec{r} - \vec{r}_i)} = \int_V \frac{k\rho(\vec{r}')dV'}{|\vec{r} - \vec{r}'|^2} \widehat{(\vec{r} - \vec{r}')} \quad (34)$$

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:

$$\vec{E} = \frac{kq}{r^2} \hat{r} \quad \Longleftrightarrow \quad d\vec{E} = \frac{k dq}{r^2} \hat{r} \quad (35)$$

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. We therefore define all of the charge densities we might need to handle these cases as:

$$\rho = \frac{dq}{dV} \iff dq = \rho dV \quad (36)$$

$$\sigma = \frac{dq}{dA} \iff dq = \rho dA \quad (37)$$

$$\lambda = \frac{dq}{d\ell} \iff dq = \rho d\ell \quad (38)$$

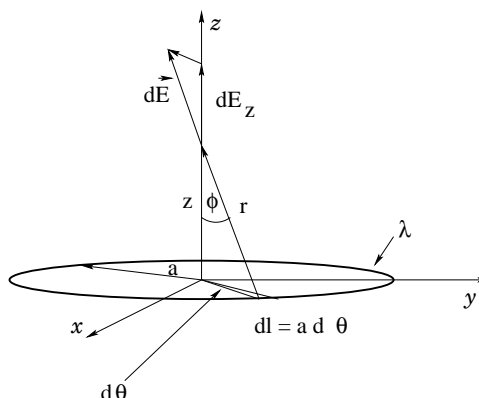
the charge per unit volume, per unit area, and per unit length respectively. In each equation I put the way we will need to *use* it – to find dq – after the defining expression.

There are thus three steps associated with solving an actual problem:

- a) Draw a picture, add a suitable coordinate system, identify the right differential chunk (one you can integrate over) and draw in the vectors needed to express $d\vec{E}$ as given above.
- b) Put down an expression for $d\vec{E}$ (or rather, usually $|d\vec{E}|$) 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{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, and especially at the beginning if you can't do the integral you won't be heavily penalized if you do the first two steps correctly. It's still something you need to work on to get the most possible credit.

Let's try some examples.

Example 2.1.1: Circular Loop of ChargeFigure 10: A charged ring with charge per unit length λ .

In figure ?? above we see a circular ring of charge of radius a and uniform charge per unit length:

$$\lambda = \frac{Q}{L} = \frac{Q}{2\pi a} \quad (39)$$

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 straightforward³².

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 θ swung around in the counterclockwise direction from the positive x -axis. The angular width of the chunk is then $d\theta$, and the length of the arc subtended is $d\ell = a d\theta$.

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:

$$dq = \lambda d\ell = \lambda a d\theta = Q \frac{d\theta}{2\pi} \quad (40)$$

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:

$$|d\vec{E}| = \frac{k_e dq}{r^2} = \frac{k_e \lambda a d\theta}{z^2 + a^2} \quad (41)$$

where we have used the pythagorean theorem to evaluate $r = \sqrt{z^2 + a^2}$ 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:

$$dE_x = -|d\vec{E}| \sin \phi \cos \theta \quad (42)$$

$$dE_y = -|d\vec{E}| \sin \phi \sin \theta \quad (43)$$

$$dE_z = |d\vec{E}| \cos \phi \quad (44)$$

³²We *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 would become very difficult – so difficult that 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.

In these equations, we must evaluate $\sin \phi$ and $\cos \phi$ using the right triangle azr :

$$\sin \phi = \frac{a}{r} = \frac{a}{(z^2 + a^2)^{1/2}} \quad (45)$$

$$\cos \phi = \frac{z}{r} = \frac{z}{(z^2 + a^2)^{1/2}} \quad (46)$$

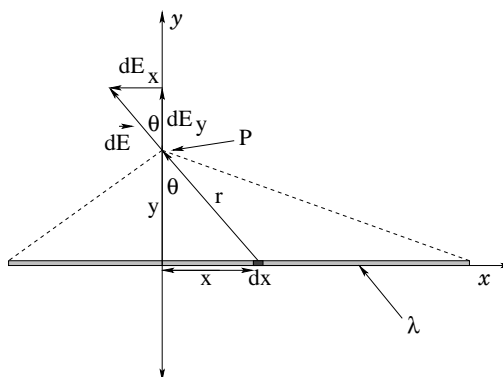
so that:

$$E_z = \int_0^{2\pi} \frac{k_e \lambda z \, a \, d\theta}{(z^2 + a^2)^{3/2}} = \frac{k_e \lambda (2\pi a) z}{(z^2 + a^2)^{3/2}} = \frac{k_e Q z}{(z^2 + a^2)^{3/2}} \quad (47)$$

Although $E_x = E_y = 0$ from symmetry as noted, it is pretty easy to actually evaluate them:

$$E_x = - \int_0^{2\pi} \frac{k_e \lambda a^2 \cos \theta \, d\theta}{(z^2 + a^2)^{3/2}} = - \frac{k_e \lambda a^2}{(z^2 + a^2)^{3/2}} \cdot \sin \theta \Big|_0^{2\pi} = 0 \quad (48)$$

(and ditto, of course, for E_y)!

Example 2.1.2: Long Straight Line of ChargeFigure 11: A straight line of charge with uniform charge per unit length λ .

In figure 11 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 and draw a y -axis through it such that P is a (shortest) distance y from the line. We pick a chunk of charge of length dx , a distance x out from the origin. The charge of our chunk is *again* given by our magic spell: “The charge of the chunk is the charge per unit length of the chunk times the length of the chunk”, or:

$$dq = \lambda dx \quad (49)$$

Finally, the magnitude of the field is given by:

$$|d\vec{E}| = \frac{k_e dq}{r^2} = \frac{k_e \lambda dx}{(x^2 + y^2)} \quad (50)$$

We need in this case to evaluate *both* dE_x and dE_y , 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 the triangles in the figure it is pretty obvious that:

$$dE_x = -|d\vec{E}| \sin \theta \quad (51)$$

$$dE_y = |d\vec{E}| \cos \theta \quad (52)$$

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 that $\cos \theta = y/r$ we get:

$$dE_y = \frac{k_e \lambda dx}{r^2} \cos \theta = \frac{k_e \lambda dx}{(x^2 + y^2)} \cos \theta = k_e \lambda y \frac{dx}{(x^2 + y^2)^{3/2}} \quad (53)$$

(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. The problem is that as we vary x , both r and θ vary as well! 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* variable. Thus:

$$x = y \tan \theta \quad (54)$$

so

$$dx = \frac{y d\theta}{\cos^2 \theta} \quad (55)$$

and

$$r = \frac{y}{\cos \theta} \quad (56)$$

If we substitute these into the expressions above we get:

$$dE_y = \frac{k_e \lambda dx}{r^2} \cos \theta = k_e \lambda \left(\frac{y d\theta}{\cos^2 \theta} \right) \left(\frac{\cos^2 \theta}{y^2} \right) \cos \theta = \frac{k_e \lambda}{y} \cos \theta d\theta \quad (57)$$

which looks *easy* to integrate! 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. Thus:

$$E_y = \frac{k_e \lambda}{y} \int_{\theta_1}^{\theta_2} \cos \theta d\theta = \frac{k_e \lambda}{y} (\sin \theta_2 - \sin \theta_1) \quad (58)$$

(where we should carefully note that θ_1 in the figure above is *negative* as drawn).

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:

$$E_x = -\frac{k_e \lambda}{y} \int_{\theta_1}^{\theta_2} \sin \theta d\theta = \frac{k_e \lambda}{y} (\cos \theta_2 - \cos \theta_1) \quad (59)$$

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:

$$E_x(\infty) = 0 \quad (60)$$

$$E_y(\infty) = \frac{2k_e \lambda}{y} \quad (61)$$

where we should recall that *every* point P has an x -coordinate in the middle of an infinite line of charge! Remember this result for later, where we will obtain it again using Gauss's Law.

Example 2.1.3: Circular Disk of Charge

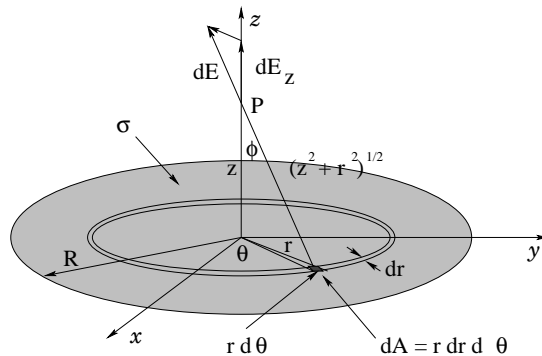


Figure 12: A charged disk with charge per unit area σ .

In figure 12 above we see a disk of charge with a uniform charge density:

$$\sigma = \frac{Q}{\pi R^2} \quad (62)$$

As before with a ring, we can only easily evaluate the field on the z -axis where we know from symmetry that $E_x = E_y = 0$. As before, we find the field of a tiny chunk of charge in suitable coordinates and sum it up using integration.

The coordinate system we choose locates the differential chunk of charge at (r, θ) inside the disk. There we mark out a small chunk of arc length $r d\theta$ as before for the ring, and of width dr , so its differential area is $dA = r d\theta dr$. As an exercise:

$$A = \int dA = \int_0^R \int_0^{2\pi} r dr d\theta = \left(\int_0^R r dr \right) \left(\int_0^{2\pi} d\theta \right) = \frac{R^2}{2} (2\pi) = \pi R^2 \quad (63)$$

and we've 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, so that a student can easily master physics if they have mastered (a rather small subset of) *one-dimensional integration methods*. They are trivially multivariate, so to speak.

At any rate, we can easily find dq from our mantra: "The charge of the chunk is the charge per unit area times the area of the chunk", or:

$$dq = \sigma dA = \sigma r dr d\theta = \frac{Q}{\pi R^2} r dr d\theta \quad (64)$$

As before, we find

$$|d\vec{E}| = \frac{k_e dq}{(r^2 + z^2)} = \frac{k_e \sigma r dr d\theta}{(r^2 + z^2)} \quad (65)$$

and

$$dE_z = |d\vec{E}| \cos \phi = \frac{k_e \sigma z r dr d\theta}{(r^2 + z^2)^{3/2}} \quad (66)$$

Finally:

$$E_z = \int dE_z = k_e \sigma z \int_0^R \int_0^{2\pi} \frac{r dr d\theta}{(r^2 + z^2)^{3/2}} \quad (67)$$

The θ integral is trivial and yields 2π . What's left is:

$$\begin{aligned} E_z &= 2\pi k_e \sigma z \int_0^R \frac{r dr}{(r^2 + z^2)^{3/2}} \\ &= \pi k_e \sigma z \int_0^R (r^2 + z^2)^{-3/2} (2r dr) \\ &= -2\pi k_e \sigma z (r^2 + z^2)^{-1/2} \Big|_0^R \\ &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2 + z^2)^{1/2}} \right) \\ &= 2\pi k_e \sigma (1 - \cos \Phi) \end{aligned} \quad (68)$$

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 $R \rightarrow \infty$ (which we can also interpret as $\Phi \rightarrow \pi/2$). In this limit, the disk of charge is *infinite* in extent – it is an infinite plane of uniform charge. The field is obviously:

$$E_z(\infty) = 2\pi k_e \sigma \quad (69)$$

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:

$$\begin{aligned}
 E_z &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2 + z^2)^{1/2}} \right) \\
 &= 2\pi k_e \sigma \left(1 - \frac{z}{z(1 + \frac{R^2}{z^2})^{1/2}} \right) \\
 &= 2\pi k_e \sigma \left(1 - (1 + \frac{R^2}{z^2})^{-1/2} \right) \\
 &\approx 2\pi k_e \sigma \left(1 - (1 - \frac{1}{2} \frac{R^2}{z^2} + \dots) \right) \\
 &\approx \pi k_e \sigma \left(\frac{R^2}{z^2} \right) \\
 &\approx \frac{k_e (\pi R^2 \sigma)}{z^2} \\
 &\approx \frac{k_e Q}{z^2}
 \end{aligned} \tag{70}$$

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 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 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; and so on. This means that we can often *guess* or very simply calculate what field of a charge distribution will look like 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.

Example 2.1.4: Advanced: Spherical Shell of Charge

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 coordinates³³ and things like the area element on the surface of a

³³Wikipedia: http://www.wikipedia.org/wiki/Spherical_Coordinate_Systems. Note well that I’m using the *physics*

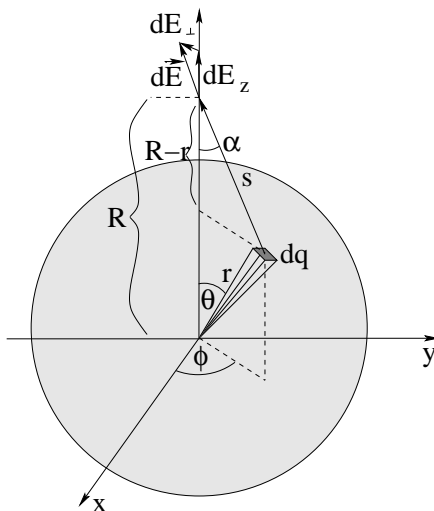


Figure 13: Geometry for finding the field of a uniform spherical **shell** of constant charge density σ by direct integration, both inside and outside. Note that θ is the angle swept *down* from the positive z axis (the equivalent of “latitude”, although measured down from the north pole and not up from 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*.

sphere:

$$dA = r^2 \sin(\theta) d\theta d\phi = -r^2 d(\cos \theta) d\phi \quad (71)$$

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:

$$\sin(\theta) d\theta = -d(\cos(\theta)) \quad (72)$$

to *change variables* from θ to $\cos(\theta)$ so that:

$$\int_0^\pi f(\cos(\theta)) \sin(\theta) d\theta = \int_{-1}^1 f(\cos(\theta)) d(\cos(\theta)) = \int_{-1}^1 f(x) dx \quad (73)$$

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 13. 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:

$$dq = \sigma dA = \sigma r^2 d(\cos \theta) d\phi \quad (74)$$

We know that the field of **just this chunk** at the point P has a magnitude:

$$dE = \frac{k_e dq}{s^2} = k_e \sigma \left(\frac{r^2 d(\cos \theta) d\phi}{s^2} \right) \quad (75)$$

Finally, we only care (for the moment, anyway) about dE_z so we might as well write it down too:

$$dE_z = dE \cos(\alpha) = k_e \sigma \left(\frac{r^2 d(\cos \theta) d\phi}{s^2} \right) \cos(\alpha) \quad (76)$$

which we can rewrite using the geometry in figure 14 as:

$$dE_z = k_e \sigma \left(\frac{r^2 d(\cos \theta) d\phi}{s^2} \right) \left(\frac{R - r \cos(\theta)}{s} \right) \quad (77)$$

convention, that is, the second of the two pictures on the right.

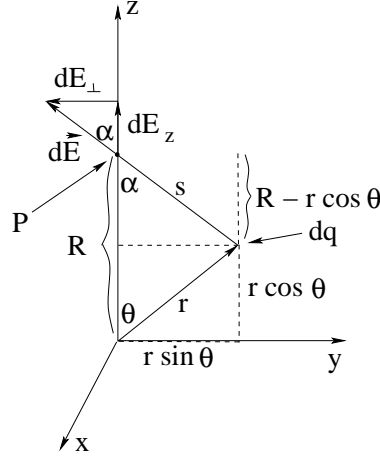


Figure 14: Geometry for the vector decomposition of $d\vec{E}$ into dE_z .

Piece of cake, right? Well, not quite. Sadly, s and $\cos(\alpha)$ depend on P , r and θ via e.g. the law of cosines³⁴ 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:

$$\begin{aligned}
 E_z &= 2\pi r^2 \sigma k_e \int_{-1}^1 \frac{(R - r \cos(\theta)) d(\cos \theta)}{s^3} \\
 &= 2\pi r^2 \sigma k_e \int_{-1}^1 \frac{(R - r \cos(\theta)) d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} \\
 &= 2\pi r^2 \sigma k_e \left\{ \int_{-1}^1 \frac{R d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} \right. \\
 &\quad \left. - \int_{-1}^1 \frac{r \cos(\theta) d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} \right\} \quad (78)
 \end{aligned}$$

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 easy) piece:

$$\begin{aligned}
 \int_{-1}^1 \frac{R d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} &= -\frac{1}{2r} \int_{-1}^1 (R^2 + r^2 - 2rR \cos(\theta))^{-3/2} (-2rR d(\cos \theta)) \\
 &= \frac{1}{r} \frac{1}{(R^2 + r^2 - 2rR \cos(\theta))^{1/2}} \Big|_{-1}^1 \\
 &= \frac{1}{r} \left\{ \frac{1}{(R^2 + r^2 - 2rR)^{1/2}} - \frac{1}{(R^2 + r^2 + 2rR)^{1/2}} \right\} \\
 &= \frac{1}{r} \left\{ \frac{1}{(R - r)} - \frac{1}{(R + r)} \right\} \\
 &= \frac{1}{r} \left\{ \frac{2r}{(R^2 - r^2)} \right\} \\
 &= \frac{2}{(R^2 - r^2)} \quad (79)
 \end{aligned}$$

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.

³⁴Wikipedia: http://www.wikipedia.org/wiki/Law_of_Cosines.

The second integral is also easy enough, at least if you remember how to **integrate by parts**:

$$\int u dv = uv - \int v du \quad (80)$$

Our chore, then, is to identify a u and a dv in the integral:

$$\int_{-1}^1 \frac{r \cos(\theta) d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} = \left(\frac{1}{-2R} \right) \int_{-1}^1 \frac{-2rR \cos(\theta) d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} \quad (81)$$

(where I've gone ahead and multiplied and divided by $-2R$, thinking ahead).

Let's let:

$$u = \cos(\theta) \quad (82)$$

and

$$\zeta = R^2 + r^2 - 2rR \cos(\theta) \quad (83)$$

so that:

$$dv = \frac{-2rR d(\cos \theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} = \zeta^{-3/2} d\zeta \quad (84)$$

We integrate this to get:

$$v = \int dv = -2\zeta^{-1/2} = \frac{-2}{(R^2 + r^2 - 2rR \cos(\theta))^{1/2}} \quad (85)$$

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.

$$\begin{aligned} \left(\frac{1}{-2R} \right) \int_{-1}^1 \frac{-2rR d(\cos \theta) \cos(\theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{3/2}} &= \left(\frac{1}{-2R} \right) \left\{ \frac{-2 \cos(\theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{1/2}} \right\} \Big|_{-1}^1 \\ &\quad - \int_{-1}^1 \frac{-2 d \cos(\theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{1/2}} \Big\} \\ &= \left(\frac{1}{-2R} \right) \left\{ \left(\frac{-2}{R-r} - \frac{2}{R+r} \right) \right. \\ &\quad \left. - \left(\frac{1}{rR} \right) \int_{-1}^1 \frac{-2rR d \cos(\theta)}{(R^2 + r^2 - 2rR \cos(\theta))^{1/2}} \right\} \\ &= \left(\frac{1}{-2R} \right) \left\{ \left(\frac{-4R}{R^2 - r^2} \right) \right. \\ &\quad \left. - \left(\frac{2}{rR} \right) (R^2 + r^2 - 2rR \cos(\theta))^{1/2} \right\} \Big|_{-1}^1 \\ &= \left(\frac{1}{-2R} \right) \left\{ \frac{-4R}{R^2 - r^2} \right. \\ &\quad \left. - \left(\frac{2}{rR} \right) [(R-r) - (R+r)] \right\} \\ &= \left(\frac{1}{-2R} \right) \left\{ \frac{-4R}{R^2 - r^2} + \frac{4}{R} \right\} \\ &= \left\{ \frac{2}{R^2 - r^2} - \frac{2}{R^2} \right\} \end{aligned} \quad (86)$$

Putting it all together we get:

$$\begin{aligned} E_z &= 2\pi r^2 \sigma k_e \left\{ \frac{2}{R^2 - r^2} - \frac{2}{R^2 - r^2} + \frac{2}{R^2} \right\} \\ &= 2\pi r^2 \sigma k_e \frac{2}{R^2} = \frac{k_e (4\pi r^2 \sigma)}{R^2} = \frac{k_e Q}{R^2} \end{aligned} \quad (87)$$

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:

$$E_z = 2\pi r^2 \sigma k_e \left\{ \int_{-1}^{R/r} \dots - \int_{R/r}^1 \right\} \quad (88)$$

(but otherwise all geometry remains the same). Ahhhh, too much work. We'll rely instead on a slightly more intuitive argument, one closely tied to **Gauss's Law**, to show that the field inside a spherical shell cancels, although (as we will see) it follows trivially from Gauss's Law itself.

What is this Gauss's Law of which I speak, you ask? Coming up next...

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*.³⁵ Maxwell's equations are, in turn, the equations of motion for the unified *dynamic* electromagnetic field, laws of nature, and 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!

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 flux of a vector field through some surface is defined to be:

$$\phi_f = \int_{\Delta S} \vec{F} \cdot \hat{n} \, dS \quad (89)$$

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{F} at all points on the surface, computes the component of \vec{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 ??, 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{n} normal to the surface is *parallel* to the electric field). ΔS is small enough

³⁵Wikipedia: [http://www.wikipedia.org/wiki/Maxwell's Equations](http://www.wikipedia.org/wiki/Maxwell's_Equations).

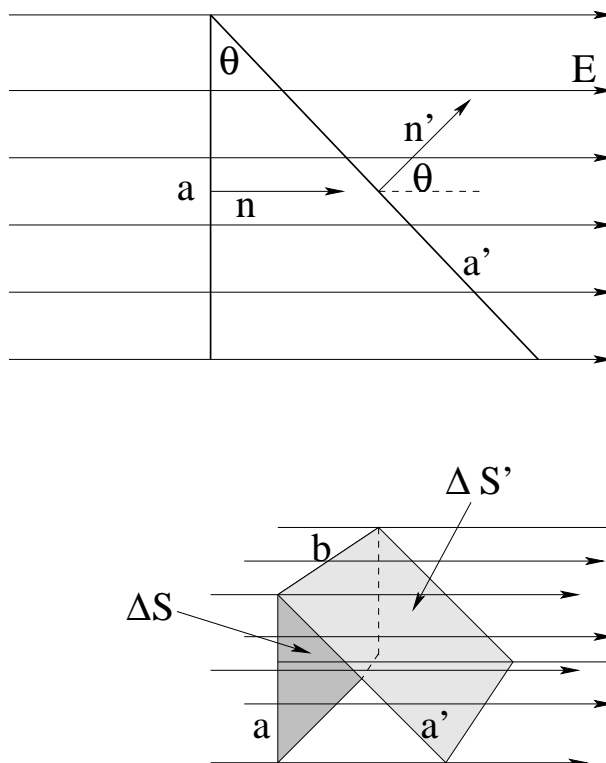


Figure 15: Geometry of the flux integral over a small surface area

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:

$$\Delta\phi_e = |\vec{E}|\Delta S = |\vec{E}|ab \quad (90)$$

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{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{E} doesn't change its *magnitude or direction* in between the two surfaces.

Note that $a = a' \cos(\theta)$, so that:

$$\Delta S' = a'b = \frac{ab}{\cos(\theta)} \quad (91)$$

If we just multiply $|\vec{E}|$ 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:

$$\begin{aligned}
 \Delta\phi'_e &= |\vec{E}| \cos(\theta) a'b \\
 &= |\vec{E}| \cos(\theta) \frac{ab}{\cos(\theta)} \\
 &= |\vec{E}| ab \\
 &= \Delta\phi_e
 \end{aligned} \tag{92}$$

We can interpret this as meaning (in words) “If \vec{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 $|\vec{E}| = \vec{E} \cdot \hat{n}$ and $|\vec{E}| \cos(\theta) = \vec{E} \cdot \hat{n}'$, so that we can write:

$$\begin{aligned}
 \lim_{\Delta S' \rightarrow 0} \Delta\phi_e &= \vec{E} \cdot \hat{n} \Delta S \\
 d\phi_e &= \vec{E} \cdot \hat{n} dS
 \end{aligned} \tag{93}$$

which *does not vary* for any possible tipping of the surface dS . The dot product precisely compensates for the increase in the area of dS as it tips relative to the direction of \vec{E} .

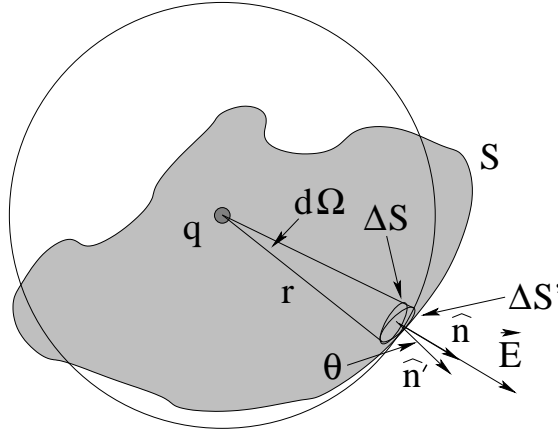


Figure 16: Point charge inside a closed surface S . Note that the flux through the tipped differential piece of the surface $\Delta S' = r^2 d\Omega / \cos \theta$ is equal to that through the *untipped* spherical piece of the surface $\Delta S = r^2 d\Omega$ that is subtended by the same solid angle $d\Omega$ and osculates the tipped surface.

Now suppose that we have a point charge surrounded by a *closed surface* S . This basically means that S is a topological deformation of a soap bubble – it *contains* a volume V with no openings. We can then imagine that the electric field of this charge is “radiated” away in all directions according to the point charge rule:

$$\vec{E} = \frac{k_e q \hat{r}}{r^2} \tag{94}$$

This situation is pictured in figure 16.

From the above, we know that if we evaluate the flux across the small patch ΔS of the *spherical* surface indicated (an osculating distance r from the charge) the field \vec{E} will be *exactly* constant and *exactly* perpendicular to that patch. In fact, the flux through that surface patch is:

$$\Delta\phi_e = \vec{E} \cdot \hat{n} \Delta S = |\vec{E}| r^2 \Delta\Omega \tag{95}$$

where $\Delta\Omega$ is the *solid angle* subtended by the cone formed by the charge and the boundary of $\Delta S = r^2 \Delta\Omega$ on the surface.

We’ve just shown that if we consider the *tipped* patch $\Delta S'$ that osculates (kisses) ΔS one end, is tipped up through an angle θ so it is actually a part of the blob shaped “arbitrary” closed surface

S' , and which subtends the *same solid angle* and hence the same “stream of flow” of the field from the charge, that the flux through it is the same:

$$\Delta\phi'_e = \vec{E} \cdot \hat{n}' \Delta S' = |\vec{E}| \cos\theta \frac{r^2 \Delta\Omega}{\cos\theta} = |\vec{E}| r^2 \Delta\Omega = \Delta\phi_e \quad (96)$$

In the differential limit, then, we can compute the flux through a small chunk of the arbitrary surface S' as:

$$\begin{aligned} d\phi_e &= \vec{E} \cdot \hat{n} dS' \\ &= |\vec{E}| r^2 d\Omega \\ &= \frac{k_e q}{r^2} r^2 d\Omega \\ &= k_e q d\Omega \end{aligned} \quad (97)$$

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. If we integrate both sides, noting that the complete solid angle (in, say, spherical polar coordinates) is:

$$\int d\Omega = \int_0^\pi \int_0^{2\pi} \sin(\theta) d\theta d\phi = 4\pi \text{ steradians} \quad (98)$$

we get:

$$\phi_e = \oint_{S'} \vec{E} \cdot \hat{n} dS = 4\pi k_e q \quad (99)$$

independent of the shape of the closed surface that we integrate over that encloses the charge q !

This is *almost* Gauss's Law. To complete our statement, we have to note first, that if the charge q is *outside* the closed surface S' , the net flux through S' is *zero*. There are a variety of ways to see this, but the easiest one is to consider S' itself to be part of a larger surface that incloses q . This creates two surfaces: one that includes the “outside” of S' and one that includes the “inside” of S' . The net flux through the two must be the same, and by changing only the sign of \hat{n} on the inner surface we can immediately see that the net flux through S' must vanish.

Second, we have to use the superposition principle. If we enclose more than one charge by S' , we just add up the fluxes so that the *total* flux is produced by the *total* charge in S' , no matter how it is distributed! Putting all this together, and getting rid of the prime on S (because it is no longer needed – the flux is the same for all closed surfaces that inclose a certain amount of charge) we get:

Gauss's Law for the Electric Field

$$\oint_{S/V} \vec{E} \cdot \hat{n} dA = 4\pi k_e \int_V \rho dV = \frac{Q_{\text{in } S}}{\epsilon_0} \quad (100)$$

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. This is the first one of *Maxwell Equation's* that we've covered so far. Only three more to go!

I used integration to compute the total charge of a continuous distribution, but of course I could equally well have summed over a bunch of discrete charges instead. The integral form will be very useful later on if you continue in physics, as it helps to transform this integral expression of Gauss's Law into a differential expression that is more useful still.

So, what's it good for? Lots! But for the moment, we'll *start* but using Gauss's law 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 planar 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.

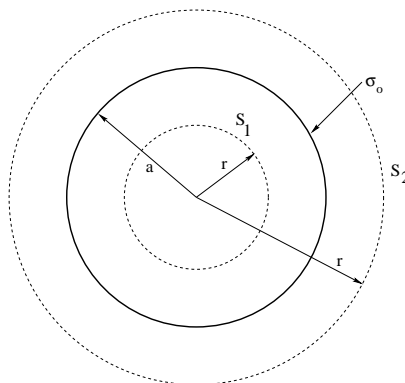


Figure 17: A spherical shell of radius a , carrying a uniform charge per unit area σ_0 . Two spherical concentric *Gaussian surfaces* S_1 (with radius $r < a$) and S_2 (with radius $r > a$) are shown.

Example 2.3.1: Spherical: A spherical shell of charge

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 17, 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{E} must point in the direction of \vec{r} and (hence) be perpendicular and constant in magnitude at all points on the Gaussian surface S_1 . Hence:

$$\phi_e = \oint_{S_1} \vec{E} \cdot \hat{r} dA = E_r \oint_{S_1} dA = E_r(4\pi r^2) \quad (101)$$

where it is presumed that everybody knows how to integrate to evaluate 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:

$$\phi_e = \oint_{S_1} \vec{E} \cdot \hat{r} dA = E_r(4\pi r^2) = \frac{Q_{S_1}}{\epsilon_0} = 0 \quad (102)$$

and solve for E_r :

$$\begin{aligned} E_r(4\pi r^2) &= 0 \\ &= \frac{0}{4\pi r^2} \\ E_r &= 0 \quad \text{for } r < a \end{aligned} \quad (103)$$

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*:

$$\phi_e = \oint_{S_2} \vec{E} \cdot \hat{r} dA = E_r \oint_{S_2} dA = E_r(4\pi r^2) \quad (104)$$

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 empty 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 $a^2 \sin(\theta) d\theta d\phi$):

$$\begin{aligned} Q_S &= \int_S \sigma_0 dA = \int_0^{2\pi} d\phi \int_0^\pi \sin \theta d\theta a^2 \sigma_0 = 2\pi a^2 \sigma_0 \int_{-1}^1 d(\cos \theta) \\ &= 4\pi a^2 \sigma_0 \end{aligned} \quad (105)$$

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 :

$$\phi_e = E_r(4\pi r^2) = \frac{Q_S}{\epsilon_0} \quad (106)$$

or

$$E_r = \frac{Q_s}{4\pi\epsilon_0 r^2} = \frac{k_e Q_s}{r^2} \quad (107)$$

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!

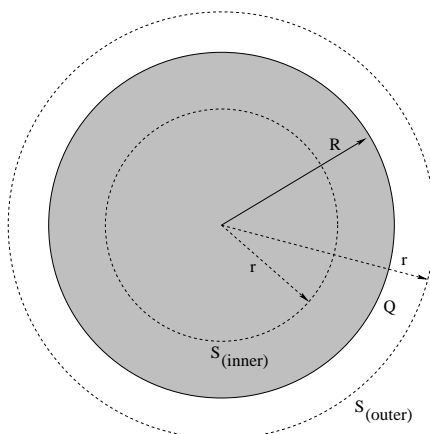
In lecture your instructor will probably do a few more difficult problems – perhaps a solid sphere of charge, or multiple spherical shells, or even a solid sphere with a charge distribution like $\rho(r) = Ar$ where A is a constant! You should be able to do *any* problem with a spherical distribution 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: 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

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.

Figure 18: A solid sphere of uniform charge density ρ and radius R .

We begin by writing Gauss's Law for the outer surface in the figure ??:

$$\begin{aligned}
 \oint_{S_{\text{outer}}} \vec{E} \cdot \hat{n} dA &= 4\pi k_e \int_{V/S} \rho dV \\
 E_r 4\pi r^2 &= 4\pi k_e \left\{ \int_0^R \int_0^{2\pi} \int_0^\pi \rho r^2 \sin(\theta) d\theta d\phi dr \right. \\
 &\quad \left. + \int_R^r 0 dV \right\} \\
 &= 4\pi k_e (2\pi \rho) \int_0^R r^2 dr \int_{-1}^1 d(\cos(\theta)) \\
 &= 4\pi k_e \left(\frac{4\pi R^3}{3} \rho \right) \\
 &= 4\pi k_e Q_{\text{total}}
 \end{aligned} \tag{108}$$

We divide both sides by $4\pi r^2$ and get:

$$E_r = \frac{k_e Q}{r^2} \quad r > R \tag{109}$$

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_{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 $\theta \rightarrow \cos(\theta)$ (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:

$$\begin{aligned}
 \oint_{S_{\text{inner}}} \vec{E} \cdot \hat{n} dA &= 4\pi k_e \int_{V/S} \rho dV \\
 E_r 4\pi r^2 &= 4\pi k_e \int_0^r \int_0^{2\pi} \int_0^\pi \rho r'^2 \sin(\theta) d\theta d\phi dr' \\
 &= 4\pi k_e (2\pi \rho) \int_0^r r'^2 dr' \int_{-1}^1 d(\cos(\theta)) \\
 &= 4\pi k_e \left(\frac{4\pi r^3}{3} \rho \right)
 \end{aligned} \tag{110}$$

We divide both sides by $4\pi r^2$ and get:

$$E_r = k_e \left(\frac{4\pi \rho r}{3} \right) \quad r < R \tag{111}$$

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,

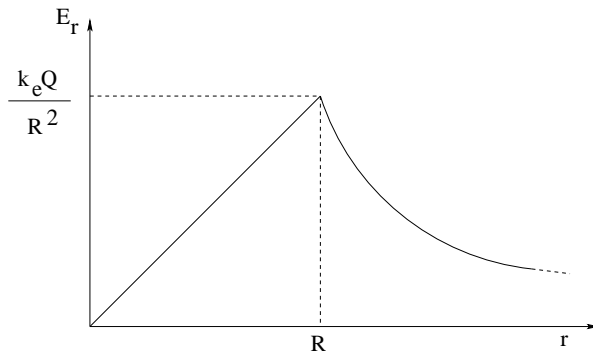


Figure 19: Electric field produced by a uniform sphere of charge both inside and outside, as a function of r .

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 at the smallest length and time scales, is far from solved³⁶.

³⁶Students 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.

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 2.3.3: Cylindrical: A cylindrical shell of charge

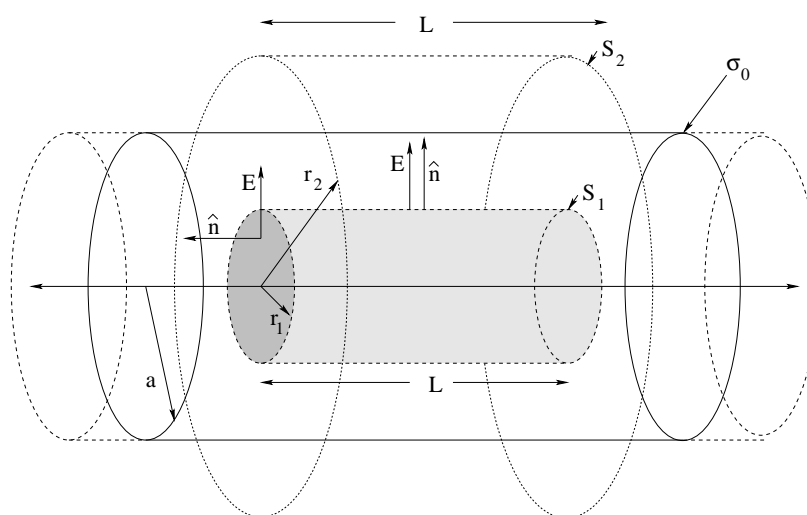


Figure 20: A cylindrical shell of radius a , carrying a uniform charge per unit area σ_0 . Two cylindrical concentric *Gaussian surfaces* S_1 (with radius $r < a$ and S_2 (with radius $r > a$) are shown.

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 field is on this surface yet, but we do know that on the *cylinder* part it must lie along \vec{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:

$$\begin{aligned}
 \phi_e &= \oint_{S_1} \vec{E} \cdot \hat{r} \, dA \\
 &= \phi_{\text{caps}} + E_r \int_{\text{Cyl}} dA \\
 &= 0 + E_r (2\pi r) L
 \end{aligned} \tag{112}$$

where it is presumed that everybody knows how to integrate to evaluate the area of a cylindrical surface of radius r and length L and knows the result³⁷. 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:

$$\phi_e = \oint_{S_1} \vec{E} \cdot \hat{r} dA = E_r(2\pi rL) = \frac{Q_{S_1}}{\epsilon_0} = 0 \quad (113)$$

and solve for E_r :

$$\begin{aligned} E_r(2\pi rL) &= 0 \\ &= \frac{0}{2\pi rL} \\ E_r &= 0 \quad \text{for } r < a \end{aligned} \quad (114)$$

We've just shown that *in general* the electric field of a cylindrical shell of charge *vanishes* inside.

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*:

$$\begin{aligned} \phi_e &= \oint_{S_2} \vec{E} \cdot \hat{r} dA \\ &= \phi_{\text{caps}} + E_r \int_C dA \\ &= E_r(2\pi r)L \end{aligned} \quad (115)$$

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 empty 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 $a d\theta dz$):

$$\begin{aligned} Q_{S_2} &= \int_S \sigma_0 dA = \int_0^{2\pi} d\theta \int_{-L/2}^{L/2} a\sigma_0 dz \\ &= 2\pi a L \sigma_0 \end{aligned} \quad (116)$$

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 :

$$\begin{aligned} \phi_e = E_r(2\pi rL) &= \frac{Q_{S_2}}{\epsilon_0} \\ E_r &= \frac{2\pi a L \sigma_0}{2\pi L \epsilon_0} \frac{1}{r} \\ &= \frac{\sigma_0 a}{\epsilon_0 r} \\ &= \frac{2k\lambda_0}{r} \end{aligned} \quad (117)$$

³⁷Think 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.

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 2.3.4: Planar: A sheet of charge

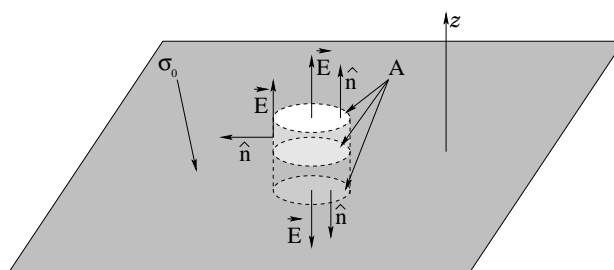


Figure 21: An (infinite) plane sheet of uniform charge per unit area σ_0 . The Gaussian surface in this case is a simple “pillbox” symmetrically drawn so it intersects the sheet as drawn.

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 \hat{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:

$$\begin{aligned}\phi_e &= \oint_S \vec{E} \cdot \hat{z} dA \\ &= \phi_{\text{side}} + 2E_z A \\ &= 2E_z A\end{aligned}\tag{118}$$

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:

$$Q_S = \int_A \sigma_0 dA = \sigma_0 A\tag{119}$$

where there really isn't much of anything to integrate or evaluate.

Finally, we write out Gauss's law and solve for E_z :

$$\begin{aligned}\phi_e = 2E_z A &= \frac{Q_S}{\epsilon_0} = \frac{\sigma_0 A}{\epsilon_0} \\ E_z &= \frac{\sigma_0}{2\epsilon_0} \\ &= 2\pi k\sigma_0\end{aligned}\tag{120}$$

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

Properties of Conductors

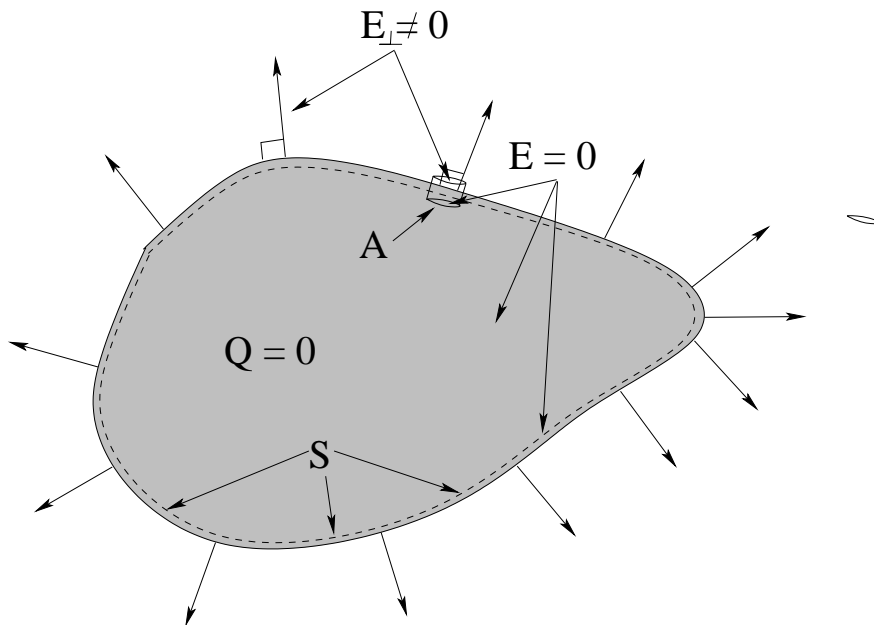


Figure 22: An arbitrary chunk of conducting material in electrostatic equilibrium can have no field inside, or else it wouldn't be in equilibrium. It can have no field tangent to its surface, or it wouldn't be in equilibrium. From these facts we can deduce several useful things about conductors in electrostatic equilibrium using Gauss's Law.

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 material known to mankind*. Consequently we have no hope of either removing all of the free electrons from a piece of metal and separating them by any appreciable macroscopic distance, or adding enough electrons so that every atom has an extra one. The material would come apart long before we succeeded.

This means that we can consider the free charge in a 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 as long as the conductor remains intact.

Now let's think a moment about the “free” bit. If we exert a force on the charges in a conductor (with, say, an electric field), they are free to move and hence will accelerate in the direction of the force. They will continue to move, speeding up, until they encounter an insulated boundary of the material, where they must stop. There they build up until they create a field of their *own* that *cancels* the applied external field, at least inside the conductor. Eventually the conductor can reach a state of *static equilibrium* where all the forces on all of the charges, including a “surface force” that holds the mobile charges inside the conductor at the surface, *cancel*.

When the conductor is in static equilibrium, we can then conclude the following:

- ***The electric field inside a conductor in static equilibrium vanishes.*** If the field were not zero, it would exert a force on the free charges inside the conductor. Since they're free, they'd move. If they move, they're not in equilibrium. So the field must be zero.
- ***The electric field parallel to the surface of the conductor in static equilibrium vanishes.*** The same argument. If there were a field component parallel to the surface, it would exert a force on charges on the surface, they can move (parallel to the surface) and hence would move, contradicting the assumption of equilibrium. Note that this does *not* restrict the field *perpendicular* to the surface of the conductor!
- ***The electric field just outside of the surface of a conductor in electrostatic equilibrium is perpendicular to the surface.*** Furthermore, from Gauss's Law we can see that it must be true that:
- $E_{\perp} = 4\pi k_e \sigma$ where σ is the charge per unit area on the surface of the conductor.

To prove this, consider a Gaussian pillbox (drawn in figure 22 above) that barely encloses the surface. Inside, the field is zero so the flux through the inside pillbox lid vanishes. The flux through the sides is zero because there is no field parallel to the sides. The flux through the *outer* pillbox surface only must therefore equal the charge inside:

$$E_{\perp} A = 4\pi k_e Q_S = 4\pi k_e \sigma A \quad (121)$$

and the result is proven.

- ***There can be no surplus charge inside a conductor in electrostatic equilibrium.*** This follows from Gauss's Law in reverse. We noted above that the field must vanish inside a conductor in equilibrium. This means that the flux through any closed surface drawn completely inside the conductor must vanish. This means in turn that the net charge inside that surface must vanish for all possible surfaces, which suffices to prove that there can be no net charge inside the conductor.

As a corollary, any unbalanced charge on a conductor in equilibrium *must be found on the surface* and must, of course, be related to \vec{E}_{\perp} at the surface.

Note well that *all of these properties are for equilibrium only!* As we will shortly learn, conductors that carry current are *not* in equilibrium 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

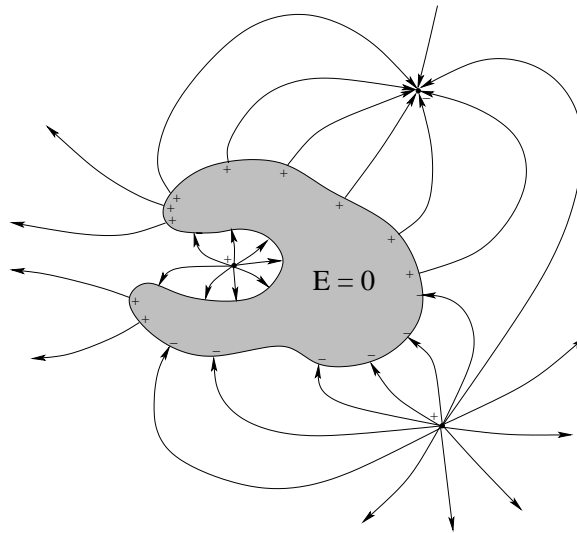


Figure 23: A conductor with an arbitrary shape near an external charge rearranges its charge into a surface charge that cancels the field inside and causes the field near the surface to be perpendicular to the surface.

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 rearranged, 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)

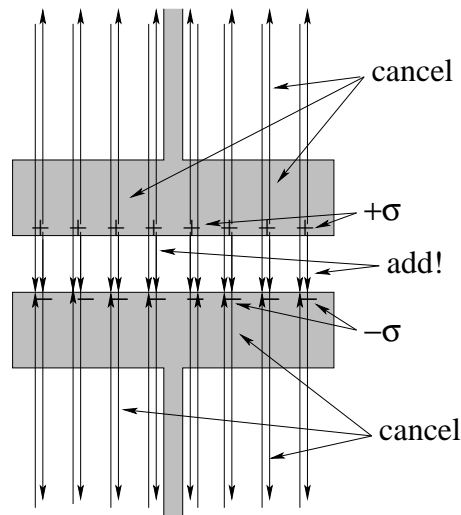


Figure 24: Opposite charges placed on two facing conducting plates spread out to form **surface charge layers**. This is exactly what is needed to **cancel the fields of the two layers in the plates themselves** while **adding together in the space between the plates**.

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 opposite 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 $E = \frac{\sigma}{2\epsilon_0} = 2\pi k_e \sigma$, 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 field lines from

each surface charge layer go *all the 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 everywhere 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:

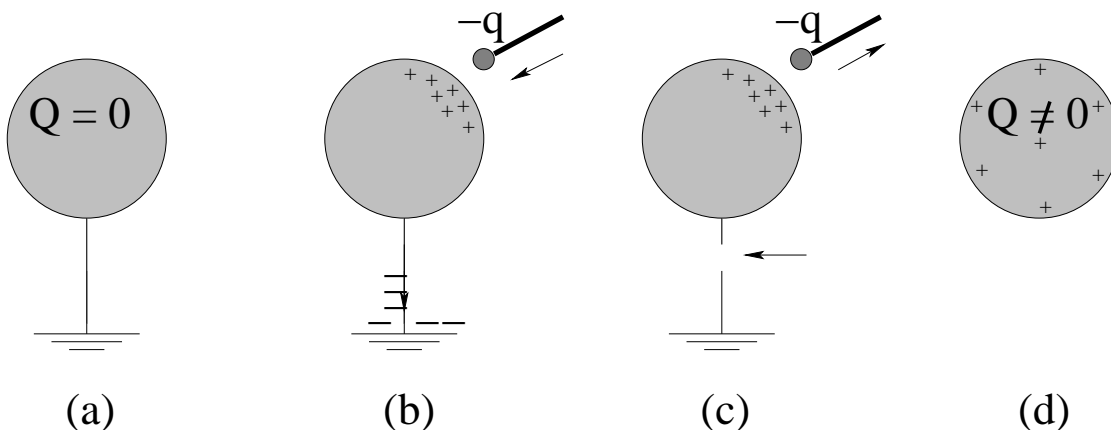


Figure 25: Charging by induction in four steps.

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.

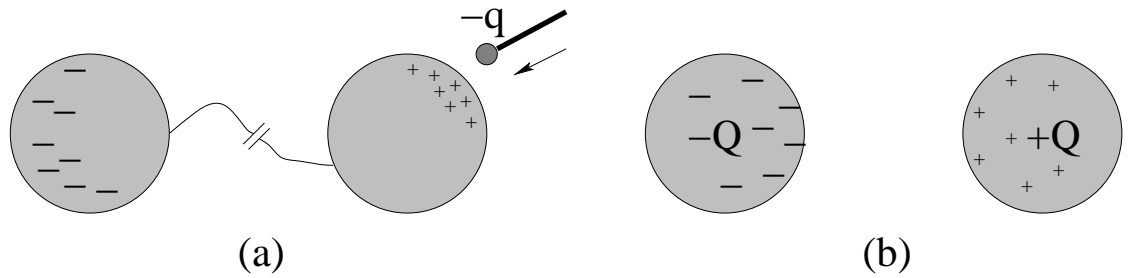


Figure 26: Charging by induction with no ground.

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 26 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 25 or 25 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.

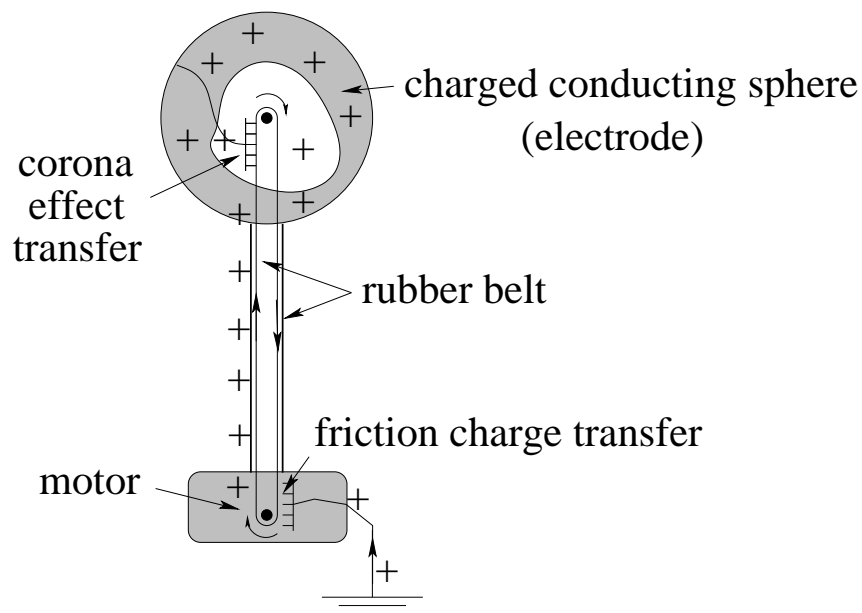


Figure 27: A Van de Graff Electrostatic Generator

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 27, which uses a flexible (rubber or silk) belt to continuously convey **triboelectrically 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 accelerators³⁸ 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.

³⁸Duke University has a high-resolution tandem Van de Graff accelerator as of the time of this writing – I helped to design its beam optics as a project in my senior year at Duke as an undergraduate.

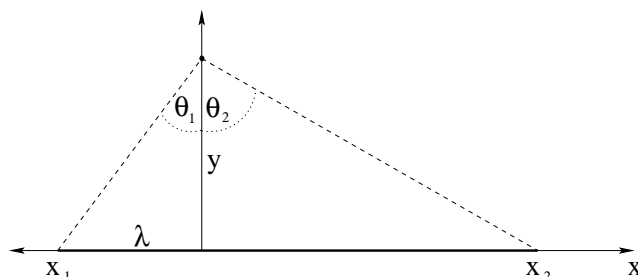
Homework for Week 2

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 (where $x_1 < x_2$ by convention) on the x axis. Find *both components* of the electric field at an arbitrary point y on the y axis. Note that x_1 and x_2 are arbitrary aside from their ordering, so your answer should make sense for e.g. $x_1 < 0$ and $x_1 > 0$.

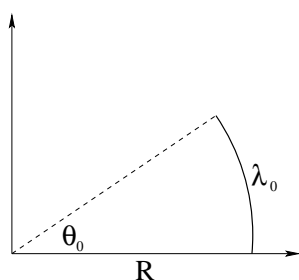
Note that this problem is *worked for you* as an *example* both in class and in the text. Why, then, you might ask yourself, is it also on the homework? Many of the examples worked in class or the text are very nearly the *only* problems of their type that can be sanely solved by e.g. integration by ordinary mortal humans. You *will not learn it* from just seeing me present it, or reading its presentation in a textbook. You must *do it yourself* – ideally enough times and carefully enough to be able to do it yourself *without looking back at the solution, easily* – in order to learn this problem and the ideas it archetypically represents and make it/them your own.

All problems that are presented in lecture, in the textbook, and as homework problems are extremely likely to show up as quiz or exam problems! In some cases, “extremely likely” means *certain*. In a subset of those cases I might even tell you that it is certain. But regardless, a *good student* will *always* be able to solve every homework problem perfectly, without looking, by exam time. An *excellent student* – one who deserves an A in the course – will be able to *explain* what they are doing as they do so (for example, to other members of their study group) and will be able to handle minor variations that make the problem not quite identical to the lecture/text/homework it is based on.

Just something to keep in mind while working on these problems in groups. The homework in this textbook is (unlike that of many textbooks) carefully designed to direct your study activities to where they will pay off. The “gold standard of learning” is being able to articulate your solutions well enough that you could teach a novice how to solve every problem of your homework a month

after doing it.

Problem 3.

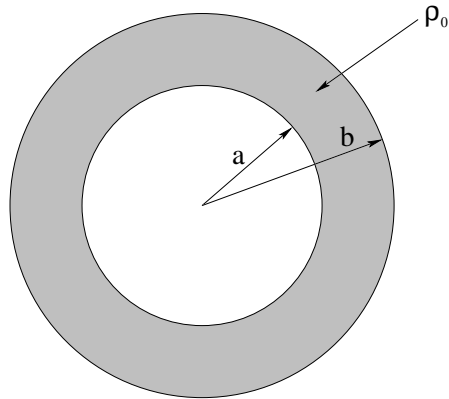


An arc of linear charge density λ_0 and radius a is centered on the origin and subtends an angle θ_0 as shown. Find the electric field at the origin.

Problem 4.

A point dipole \vec{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 you may want to look back at some of your homework problems from last week as you do this. Which ones are likely to help you out? How do we make a “point dipole”?

Problem 5.

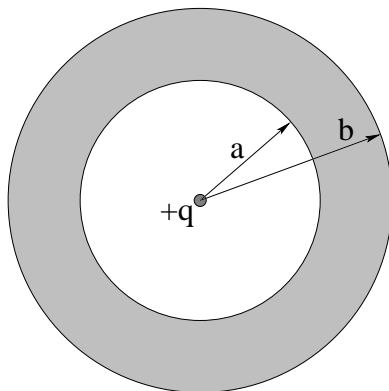
A thick, nonconducting spherical shell of inner radius a and outer radius b has a uniform volume charge density $\rho(r) = \rho_0$.

- a) Find the total charge of the shell.
- b) Find the electric field everywhere.

Problem 6.

An infinitely long nonconducting cylindrical *shell* of inner radius a and outer radius b carries a uniform volume charge density $\rho(r) = \rho_0$.

- a) Find the electric field everywhere.
- b) Let $a = 0$. Find the electric field (now that of a uniform cylinder of charge) everywhere.

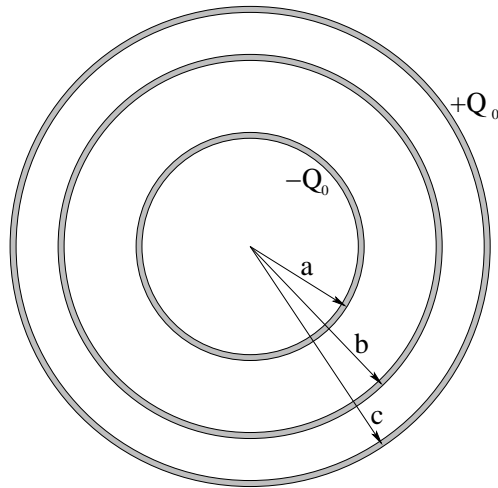
Problem 7.

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.

- a) Use Gauss's Law and the properties of conductors in equilibrium to find the electric field in the regions $r < a$, $a < r < b$, $b < r$.
- b) Find the charge density on the inner and outer surfaces of the shell.

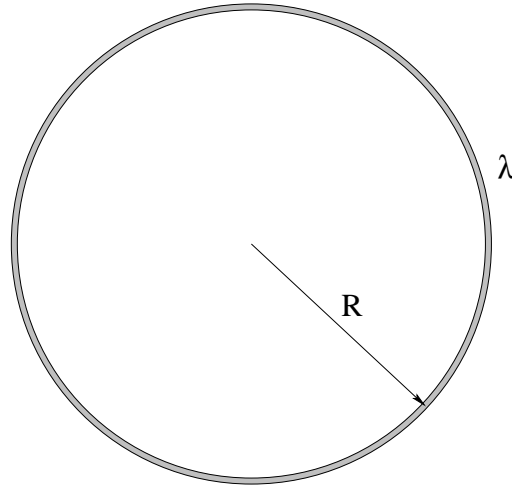
Problem 8.

A conducting neutral sphere of radius R is placed in a uniform electric field $\vec{E} = E_0 \hat{z}$. 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.

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 10.

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.

- 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.
- Does this field point towards the nearest arc of the ring or the farthest arc of the ring?
- Suppose a charge $-q$ is placed at the center of the ring (at equilibrium). Is this equilibrium stable³⁹?
- 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 too⁴⁰?

Problem 11.

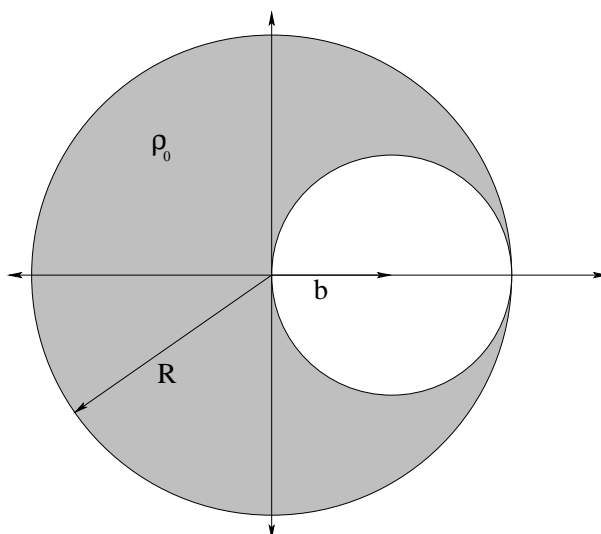
A uniformly charged nonconducting sphere of radius a is centered on the origin and has a uniform charge density $\rho(r) = \rho_0$.

- Show that at a point within the sphere a distance r from the center the electric field is given by:

$$\vec{E} = \frac{\rho_0 \vec{r}}{3\epsilon_0} = \frac{4\pi k \rho_0 \vec{r}}{3}$$

³⁹As 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.

⁴⁰Alternatively, 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.

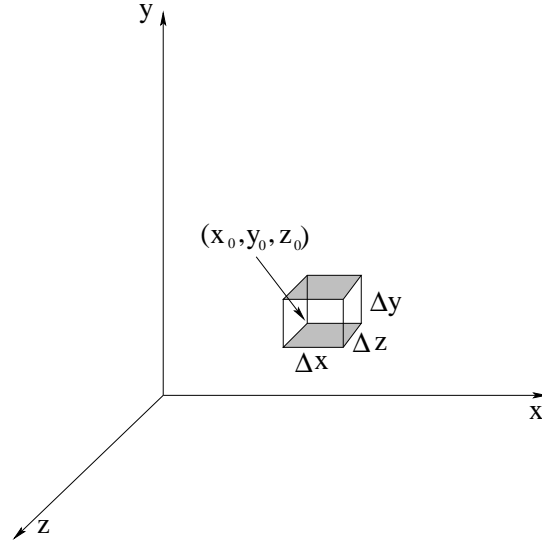


- b) Material is removed from the sphere to create a spherical cavity of radius $b = a/2$ with center at $x = b$ on the x axis (shown above). Show that the electric field inside the cavity is *uniform* and equal to:

$$\vec{E} = \frac{\rho_0 \vec{b}}{3\epsilon_0} = \frac{4\pi k \rho_0 \vec{b}}{3}$$

in magnitude (where $\vec{b} = b\hat{x}$).

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 12.

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:

$$\begin{aligned}
 \vec{E}(\vec{r}_0 + \Delta\vec{r}) = & \left(E_x(\vec{r}_0) + \Delta x \left. \frac{\partial E_x}{\partial x} \right|_{\vec{r}_0} + \Delta y \left. \frac{\partial E_x}{\partial y} \right|_{\vec{r}_0} + \right. \\
 & \left. \Delta z \left. \frac{\partial E_x}{\partial z} \right|_{\vec{r}_0} + \dots \right) \hat{x} + \\
 & \left(E_y(\vec{r}_0) + \Delta x \left. \frac{\partial E_y}{\partial x} \right|_{\vec{r}_0} + \Delta y \left. \frac{\partial E_y}{\partial y} \right|_{\vec{r}_0} + \right. \\
 & \left. \Delta z \left. \frac{\partial E_y}{\partial z} \right|_{\vec{r}_0} + \dots \right) \hat{y} + \\
 & \left(E_z(\vec{r}_0) + \Delta x \left. \frac{\partial E_z}{\partial x} \right|_{\vec{r}_0} + \Delta y \left. \frac{\partial E_z}{\partial y} \right|_{\vec{r}_0} + \right. \\
 & \left. \Delta z \left. \frac{\partial E_z}{\partial z} \right|_{\vec{r}_0} + \dots \right) \hat{z} +
 \end{aligned} \tag{122}$$

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:

$$\sum_{\text{sides}} \vec{E} \cdot \hat{n} \Delta A = \phi_{\text{net}} = \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right) \Delta V = \vec{\nabla} \cdot \vec{E} \Delta V$$

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: $\Delta x \rightarrow dx$, etc. Then Gauss's Law as we have thus far learned it becomes the following vector differential form:

$$\sum_{\text{sides}} \vec{E} \cdot \hat{n} dA = \vec{\nabla} \cdot \vec{E} dV = \frac{\rho}{\epsilon_0} dV$$

or

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \quad (123)$$

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:

$$\Delta U(\vec{x}_0 \rightarrow \vec{x}_1) = - \int_{\vec{x}_0}^{\vec{x}_1} \vec{F} \cdot d\vec{x} \quad (124)$$

where \vec{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*:

$$\vec{F} = -\vec{\nabla} U \quad (125)$$

- 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 ∞ :

$$U(\vec{x}) = - \int_{\infty}^{\vec{x}} \vec{F} \cdot d\vec{x} \quad (126)$$

For point charges q_1 and q_2 , it is just:

$$U(\vec{x}_1, \vec{x}_2) = \frac{kq_1q_2}{|\vec{x}_1 - \vec{x}_2|} \quad (127)$$

- Since the potential energy is just a scalar and satisfies the superposition principle, we can evaluate the total energy of a system of point charges as:

$$U_{\text{tot}} = \frac{1}{2} \sum_{i \neq j} \frac{kq_iq_j}{|\vec{x}_i - \vec{x}_j|} \quad (128)$$

(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:

$$V(\vec{x}) = \lim_{q_0 \rightarrow 0} \frac{U(\vec{x})}{q_0} \quad (129)$$

so that the potential of a point charge in coordinates centered on the charge is just:

$$V(\vec{r}) = \frac{kq}{r} \quad (130)$$

- The potential is to the field as the potential energy is to the force, so:

$$V(\vec{x}) = - \int \vec{E} \cdot d\vec{x} + V_0 \quad (131)$$

with V_0 and arbitrary constant of integration, used to set a suitable zero of the potential energy. For compact charge distributions:

$$V(\vec{x}) = - \int_{\infty}^{\vec{x}} \vec{E} \cdot d\vec{x} \quad (132)$$

and

$$\vec{E} = -\vec{\nabla}V \quad (133)$$

- The potential of a charge distribution can obviously be evaluated by superposition:

$$V_{\text{tot}}(\vec{x}) = \sum_i \frac{kq_i}{|\vec{x} - \vec{x}_i|} \quad (134)$$

or

$$V_{\text{tot}}(\vec{x}) = \int \frac{k dq_0}{|\vec{x} - \vec{x}_0|} = \int \frac{k \rho(\vec{x}_0) d^3 r_0}{|\vec{x} - \vec{x}_0|} \quad (135)$$

- 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:

$$\Delta U(\vec{x}_0 \rightarrow \vec{x}_1) = - \int_{\vec{x}_0}^{\vec{x}_1} \vec{F} \cdot d\vec{x} \quad (136)$$

The corresponding relation between the potential energy thus defined and the force is (as usual):

$$\vec{F} = -\vec{\nabla}U \quad (137)$$

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*:

$$\Delta U(\vec{x}) = - \int \vec{F} \cdot d\vec{x} + U_0 \quad (138)$$

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:

$$\Delta U(\vec{x}) = - \int_{\infty}^{\vec{x}} \vec{F} \cdot d\vec{x} \quad (139)$$

(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:

$$U = \frac{kq_1q_2}{|\vec{x}_1 - \vec{x}_2|} \quad (140)$$

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*:

$$U_{\text{tot}} = \frac{1}{2} \sum_{i \neq j} \frac{kq_iq_j}{|\vec{x}_i - \vec{x}_j|} \quad (141)$$

Note that in this sum the $1 \rightarrow 2$ 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:

$$U_{\text{tot}} = \sum_{i < j} \frac{kq_iq_j}{|\vec{x}_i - \vec{x}_j|} \quad (142)$$

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:

$$V(\vec{x}) = \lim_{q_0 \rightarrow 0} \frac{U(\vec{x})}{q_0} \quad (143)$$

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:

$$1 \text{ Volt} = \frac{1 \text{ Joule}}{1 \text{ Coulomb}} \quad (144)$$

If we apply this rule compute the potential at \vec{x} produced by a point charge q at the origin of coordinates, we get:

$$V(\vec{x}) = \lim_{q_0 \rightarrow 0} \frac{1}{q_0} \frac{kqq_0}{|\vec{x} - 0|} = \frac{kq}{r} \quad (145)$$

where $r = |\vec{x}|$. Alternatively we could use the definition of the field relative to the force to define:

$$V(\vec{x}) = - \int \vec{E} \cdot d\vec{x} + V_0 \quad (146)$$

For charge distributions with compact support, we by convention pick the zero of potential at ∞ so that:

$$V(\vec{x}) = - \int_{\infty}^{\vec{x}} \vec{E} \cdot d\vec{x} \quad (147)$$

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:

$$\Delta V(\vec{x}_1 \rightarrow \vec{x}_2) = - \int_{\vec{x}_1}^{\vec{x}_2} \vec{E} \cdot d\vec{x} \quad (148)$$

We can invert these relations to obtain:

$$\vec{E} = -\vec{\nabla}V \quad (149)$$

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:

$$U = qV \quad (150)$$

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:

$$V_{\text{tot}}(\vec{x}) = \sum_i \frac{kq_i}{|\vec{x} - \vec{x}_i|} \quad (151)$$

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:

$$\begin{aligned} V_{\text{tot}}(\vec{x}) &= \int \frac{k dq_0}{|\vec{x} - \vec{x}_0|} \\ &= \int \frac{k \rho(\vec{x}_0) d^3r_0}{|\vec{x} - \vec{x}_0|} \quad \text{volume} \end{aligned} \quad (152)$$

$$= \int \frac{k \sigma(\vec{x}_0) d^2r_0}{|\vec{x} - \vec{x}_0|} \quad \text{area} \quad (153)$$

$$= \int \frac{k \lambda(\vec{x}_0) dr_0}{|\vec{x} - \vec{x}_0|} \quad \text{line} \quad (154)$$

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

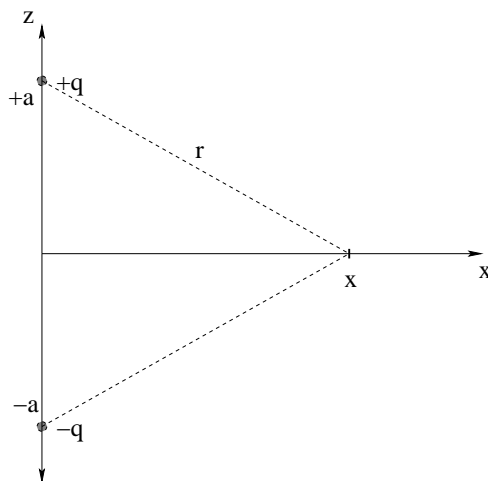


Figure 28: A simple dipole aligned with the z -axis.

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:

$$\begin{aligned} V(x) &= \sum_{i=1}^2 \frac{k_e q_i}{r_i} \\ &= \frac{k_e q}{(x^2 + a^2)^{1/2}} - \frac{k_e q}{(x^2 + a^2)^{1/2}} = 0 \end{aligned} \quad (155)$$

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 $\vec{E} = -\vec{\nabla}V$. We are ever so tempted to go e.g.:

$$E_z = -\frac{dV}{dz} = -\frac{d0}{dz} = 0 \quad (156)$$

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.

$$V(x, z) = \frac{k_e q}{(x^2 + (a - z)^2)^{1/2}} - \frac{k_e q}{(x^2 + (a + z)^2)^{1/2}} \quad (157)$$

Now we can differentiate:

$$\begin{aligned}
 E_z &= -\frac{d}{dz} \frac{k_e q}{(x^2 + (a - z)^2)^{1/2}} + \frac{d}{dz} \frac{k_e q}{(x^2 + (a + z)^2)^{1/2}} \\
 &= -\frac{k_e q(a - z)}{(x^2 + (a - z)^2)^{3/2}} - \frac{k_e q(a + z)}{(x^2 + (a + z)^2)^{1/2}}
 \end{aligned} \tag{158}$$

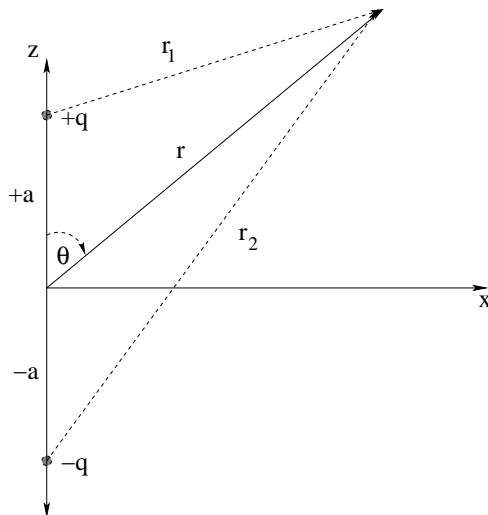
NOW we can let $z \rightarrow 0$ to find out what the field is on the x -axis (adding and cancelling terms as necessary, and substituting $p_z = 2qa$ in for the dipole moment):

$$\begin{aligned}
 E_z &= -\frac{2k_e qa}{(x^2 + a^2)^{3/2}} \\
 &= -\frac{k_e p_z}{(x^2 + a^2)^{3/2}}
 \end{aligned} \tag{159}$$

Compare this to equation (23)! Hmmm, looks the same⁴¹! 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).

⁴¹Allowing, 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 SpaceFigure 29: 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*:

$$r_1 = +\sqrt{r^2 + a^2 - 2ar \cos(\theta)} \quad (160)$$

$$r_2 = +\sqrt{r^2 + a^2 + 2ar \cos(\theta)} \quad (161)$$

so that the potential can be read off by inspection:

$$V(r, \theta) = \frac{k_e q}{(r^2 + a^2 - 2ar \cos(\theta))^{1/2}} - \frac{k_e q}{(r^2 + a^2 + 2ar \cos(\theta))^{1/2}} \quad (162)$$

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:

$$\begin{aligned}
 \lim_{r \gg a} V(r, \theta) &= \frac{k_e q}{(r^2 + a^2 - 2ar \cos(\theta))^{1/2}} - \frac{k_e q}{(r^2 + a^2 + 2ar \cos(\theta))^{1/2}} \\
 &= \frac{k_e q}{r} \left\{ \left(1 - 2\frac{a}{r} \cos(\theta) + \frac{a^2}{r^2}\right)^{-1/2} - \left(1 + 2\frac{a}{r} \cos(\theta) + \frac{a^2}{r^2}\right)^{-1/2} \right\} \\
 &= \frac{k_e q}{r} \left\{ \left(1 + \frac{a}{r} \cos(\theta) - \frac{a^2}{2r^2} + \dots\right) - \left(1 - \frac{a}{r} \cos(\theta) - \frac{a^2}{2r^2} + \dots\right) \right\} \\
 &= \frac{k_e q}{r} \left\{ 2\frac{a}{r} \cos(\theta) + \mathcal{O}\left(\frac{a^3}{r^3}\right) \right\} \\
 &\approx \frac{k_e 2qa}{r^2} \cos(\theta) \\
 &\approx \frac{k_e p_z}{r^2} \cos(\theta) = k_e \frac{\vec{p} \cdot \hat{r}}{r^2}
 \end{aligned} \tag{163}$$

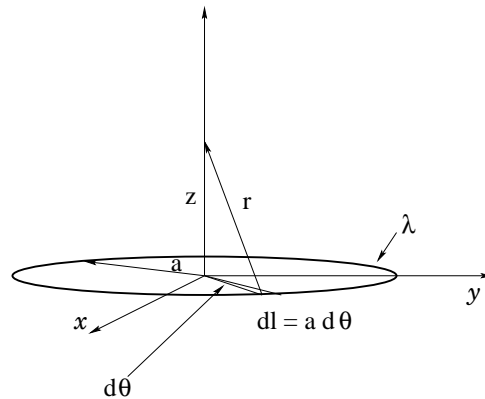
where \hat{r} is a unit vector in the \vec{r} direction. (We used our freedom to rotate the coordinate system so that \vec{p} points in an arbitrary direction instead of \vec{z} to guess the last result.)

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 chargeFigure 30: 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:

- In suitable coordinates, define a differential “chunk” of the charge. In this problem, that would be a differential-size arc segment of the ring.
- 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).
- 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!

- 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).
- Simplify, if appropriate or required.
- 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.

$$dl = a d\theta \quad (164)$$

defines a differential chunk of the ring. Its charge is:

$$dq = \lambda dl \quad (165)$$

The differential potential of this chunk at a point on the z -axis is:

$$dV(z) = \frac{k_e dq}{r} = \frac{k_e \lambda a d\theta}{(z^2 + a^2)^{1/2}} \quad (166)$$

We integrate over all of the chunks of charge that make up the ring by integrating θ from 0 to 2π :

$$\begin{aligned} V(z) = \int dV &= \int_0^{2\pi} \frac{k_e \lambda a d\theta}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e (2\pi a) \lambda}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e Q}{r} \end{aligned} \quad (167)$$

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:

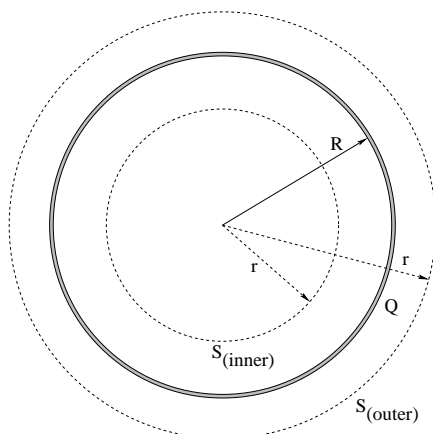
$$\begin{aligned} E_z &= -\frac{d}{dz} \frac{k_e (2\pi a) \lambda}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e (2\pi a) \lambda z}{(z^2 + a^2)^{3/2}} \\ &= \frac{k_e Q z}{(z^2 + a^2)^{3/2}} \end{aligned} \quad (168)$$

Compare this to equation (47) 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.

Figure 31: A spherical shell of charge of radius R .

Example 3.4.4: Potential of a Spherical Shell of Charge

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:

$$\begin{aligned}\vec{E} &= 0 & (r < R) \\ \vec{E} &= \frac{k_e Q}{r^2} \hat{r} & (r > R)\end{aligned}$$

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:

$$V(\vec{r}) = - \int_{\infty}^{\vec{r}} \vec{E} \cdot d\vec{l} \quad (169)$$

In this case, we integrate piecewise from the outside in to find the field outside and inside of the sphere, accordingly. Outside:

$$V(\vec{r}) = - \int_{\infty}^r \frac{k_e Q}{r^2} dr = \frac{k_e Q}{r} \quad (170)$$

for all $r > R$. Inside:

$$V(\vec{r}) = - \int_{\infty}^R \frac{k_e Q}{r^2} dr - \int_R^r 0 dr = \frac{k_e Q}{R} \quad (171)$$

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*.

Example 3.4.5: Advanced: Spherical Shell of Charge

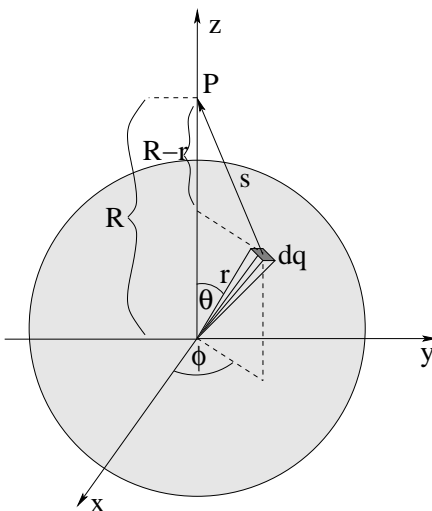


Figure 32: Geometry for finding the potential of a uniform spherical **shell** of constant charge density σ by direct integration.

Consider figure 32. 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 :

$$dV = \frac{k_e dq}{s} = \frac{k_e \sigma r^2 d\cos(\theta) d\phi}{(R^2 + r^2 - 2Rr \cos(\theta))^{1/2}} \quad (172)$$

We integrate both sides, the right hand side over the entire solid angle:

$$V = \int dV = \int \frac{k_e dq}{s} = \int_{-1}^1 \int_0^{2\pi} \frac{k_e \sigma r^2 d \cos(\theta) d\phi}{(R^2 + r^2 - 2Rr \cos(\theta))^{1/2}} \quad (173)$$

We can do the ϕ integral immediately and factor out all the constants:

$$V = 2\pi r^2 \sigma k_e \int_{-1}^1 \frac{d \cos(\theta)}{(R^2 + r^2 - 2Rr \cos(\theta))^{1/2}} \quad (174)$$

This is much easier to integrate than the vector relation of the field chapter example:

$$\begin{aligned} V &= 2\pi r^2 \sigma k_e \int_{-1}^1 \frac{d \cos(\theta)}{(R^2 + r^2 - 2Rr \cos(\theta))^{1/2}} \\ &= \frac{2\pi r^2 \sigma k_e}{-2Rr} \int_{-1}^1 \frac{-2Rr d \cos(\theta)}{(R^2 + r^2 - 2Rr \cos(\theta))^{1/2}} \\ &= \frac{2\pi r^2 \sigma k_e}{-2Rr} 2 (R^2 + r^2 - 2Rr \cos(\theta))^{1/2} \Big|_{-1}^1 \\ &= \frac{2\pi r^2 \sigma k_e}{-2Rr} 2 ((R - r) - (R + r)) \\ &= \frac{2\pi r^2 \sigma k_e}{-2Rr} (-2r) \\ &= \frac{k_e (4\pi r^2 \sigma)}{R} = \frac{k_e Q}{R} \end{aligned} \quad (175)$$

Much, much easier!

Example 3.4.6: Potential of a Uniform Ball of Charge

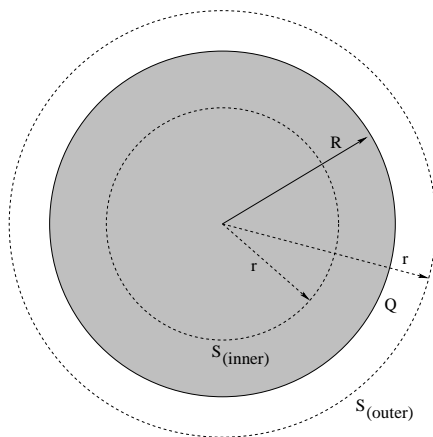


Figure 33: A solid sphere of uniform charge density ρ and radius R .

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:

$$E_r = \frac{k_e \left(\frac{4\pi R^3 \rho}{3} \right)}{r^2} = \frac{k_e Q}{r^2} \quad r > R$$

and

$$E_r = k_e \left(\frac{4\pi\rho}{3} \right) r = \frac{\rho r}{3\epsilon_0} \quad r < R$$

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):

$$\begin{aligned} V(r) &= - \int_{\infty}^r \vec{E} \cdot d\vec{l} = - \int_{\infty}^r E_r > R dr \\ &= - \int_{\infty}^r k_e Q r'^{-2} dr' \\ &= \frac{k_e Q}{r} \quad r > R \end{aligned} \tag{176}$$

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:

$$\begin{aligned} V(r) &= - \int_{\infty}^r \vec{E} \cdot d\vec{l} = - \int_{\infty}^R E_r > R dr - \int_R^r E_r < R dr \\ &= - \int_{\infty}^R k_e \left(\frac{4\pi R^3 \rho}{3} \right) r'^{-2} dr' - \int_R^r k_e \left(\frac{4\pi \rho}{3} \right) r' dr' \\ &= k_e \left(\frac{4\pi R^3 \rho}{3} \right) + k_e \left(\frac{2\pi \rho}{3} \right) \{R^2 - r^2\} \\ &= 2\pi k_e \rho R^2 - k_e \left(\frac{2\pi \rho}{3} \right) r^2 \quad r < R \end{aligned} \tag{177}$$

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{E} is continuous but not smooth (look back at figure ?? 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!

There is one more thing that we need to do before abandoning the ball of charge. Suppose we are handed such a ball. A perfectly reasonable question for any physics groupie is “How much work did it take to assemble all of this charge?” After all, the charge is mutually repulsive – every bit of charge we put into the ball had to be brought in “from infinity” against the field of the charge that is already there. This latter insight is the key to writing down a simple integral to tell us how much work was done, and hence *what the potential energy of a uniform ball of charge is*.

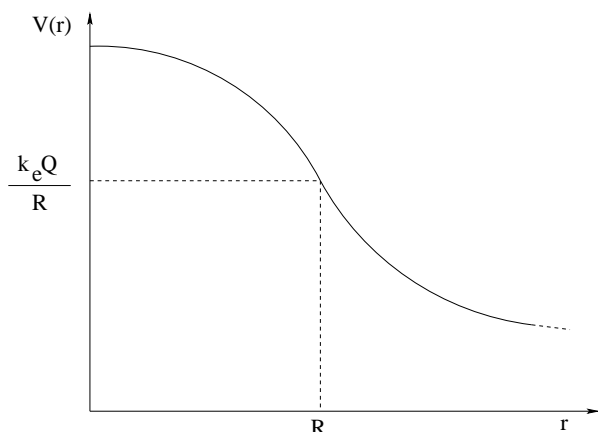


Figure 34: The potential produced by a uniform sphere of charge both inside and outside, as a function of r .

Suppose we have built a ball of radius r and total charge:

$$Q(r) = \frac{4\pi}{3} \rho r^3 \quad (178)$$

(so far). We know (or can figure out easily given the results just above) that the potential on its surface is just:

$$V(r) = \frac{k_e Q(r)}{r} = \frac{4\pi k_e}{3} \rho r^2. \quad (179)$$

Now imagine bringing in a differential chunk of charge dQ and spreading it around on the surface, increasing the radius of the ball just a bit. The work *we* have to do bringing the charge from ∞ to the surface of the ball (which is also the increase in the potential energy of the ball) is:

$$dW_{(\text{us})} = dU_{(\text{ball})} = V(r)dQ = V(r)\rho 4\pi r^2 dr \quad (180)$$

where we use the fact that the charge of a thin shell of radius r and thickness dr is just the volume of the shell times the charge per unit volume. We can now add up this increment of energy by integrating to “build a ball”:

$$\begin{aligned} U &= \int_0^R V(r)\rho 4\pi r^2 dr \\ &= \int_0^R \frac{4\pi k_e}{3} \rho r^2 \rho 4\pi r^2 dr \\ &= k_e \frac{16\pi^2 \rho^2}{3} \int_0^R r^4 dr \\ &= k_e \frac{16\pi^2 \rho^2}{3} \frac{R^5}{5} \\ &= \frac{3}{5} \frac{k_e \left(\frac{4\pi R^3}{3} \rho \right)^2}{R} = \frac{3}{5} \frac{k_e Q^2}{R} = \frac{3}{5} V(R)Q \end{aligned} \quad (181)$$

This is an extremely interesting result. Note first that if we knew *nothing* about how the charge was distributed and were asked to estimate its energy, the only sensible answer we can give (that makes dimensional sense) is $U = V \times Q$. Charge times potential equals potential energy. Of course we don’t expect the energy to be *exactly* this – we expect it to be less, because we can bring in the first bits of charge “for free” and do ever more work as we build up the ball – we expect it to be something *less* than this estimate.

Later we’ll do more examples of this sort of integral when we discuss capacitance, and will find that the *form* of this result is quite general, but (as one might expect) the leading fraction will vary

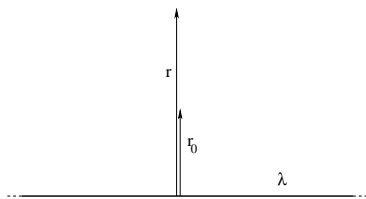
depending on the *details* of how the charge we assemble is distributed. For a conducting sphere (where all the charge resides on the outside) or spherical shell of charge, for example, it will be $1/2$. See if you can show this.

As a final note of interest, observe how the potential energy of the ball of charge scales with its radius! As any fixed amount of charge is compressed into smaller and smaller balls so that $R \rightarrow 0$, we see that $U(R \rightarrow 0) \rightarrow \infty$! 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 and certain properties of electrons.

To compute the first, assume that $Q = +e$ and $R = 10^{-15}$ meters (one fermi) – a model for the *proton* as a ball of charge. If one computes $k_e e/R$ for this in SI units (Volts) and multiplies by the remaining $+e$ to get $k_e e^2/R$ in eV, one gets $+1.44$ MeV – the *order of magnitude* of the energy bound up in the electrostatic field of the charge of a proton. Nuclear forces that glue all of this charge together (with *gluons*) 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 order of ten million times as much energy per interaction than does e.g. burning one carbon atom into carbon dioxide.

The second requires 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”⁴². 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 $U = 0.5$ 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.

⁴²Wikipedia: http://www.wikipedia.org/wiki/Classical_Electron_Radius.

Example 3.4.7: Potential of an Infinite Line of ChargeFigure 35: 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:

$$\vec{E} = \frac{2k_e\lambda}{r}\hat{r}$$

in cylindrical coordinates, so that \hat{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:

$$V(r) = - \int_{\infty}^r \frac{2k_e\lambda}{r'} dr' = -2k_e\lambda (\ln(r) - \ln(\infty)) = \infty - 2k_e\lambda \ln(r) \quad (182)$$

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 :

$$V(r) = - \int_{r_0}^r \frac{2k_e\lambda}{r'} dr' = -2k_e\lambda (\ln(r) - \ln(r_0)) = -2k_e\lambda \ln\left(\frac{r}{r_0}\right) \quad (183)$$

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:

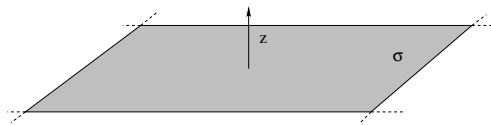
$$V(r) = -2k_e\lambda \ln(r) \quad (184)$$

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:

$$\Delta V(r_1 \rightarrow r_2) = - \int_{r_1}^{r_2} \frac{2k_e\lambda}{r'} dr' = 2k_e\lambda \ln\left(\frac{r_1}{r_2}\right) \quad (185)$$

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 $\Delta V(r_1 \rightarrow r_2) = V(r_2) - V(r_1)$ 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).

Figure 36: An “infinite” plane of uniform charge density σ .

Potential of an Infinite Plane of Charge

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:

$$V(z) = - \int_{\infty}^z 2\pi k_e \sigma \, dz = \infty - 2\pi k_e \sigma z \quad (186)$$

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!

$$V(z) = - \int_0^z 2\pi k_e \sigma \, dz = 0 - 2\pi k_e \sigma z = -2\pi k_e \sigma z \quad (187)$$

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: 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 $\vec{E} = 0$ inside a conductor in electrostatic equilibrium and that $\vec{E}_{||} = 0$ 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:

$$\Delta V = - \int_{\vec{x}_0}^{\vec{x}_1} \vec{E} \cdot d\vec{x} = 0 \quad (188)$$

Note that this doesn't mean that the potential of the conductor is *zero*, only that it is a *constant*. That is consistent:

$$\vec{E} = -\vec{\nabla}V_0 = 0 \quad (189)$$

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 arrangement. In the next chapter this will be the basis of the notion of *capacitance*.

Charge Sharing

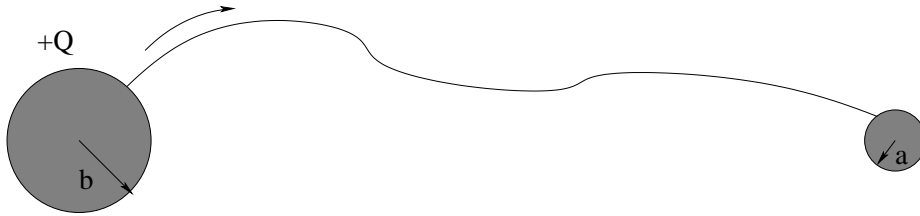


Figure 37: Charge sharing between two distant conductors connected by a wire. They become equipotential, with charge transferred (shared) between them to make it so.

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 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:

$$V_b = - \int_{\infty}^b E_r dr = \frac{kQ}{b} \quad (190)$$

everywhere inside sphere b while

$$V_a = 0 \quad (191)$$

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:

$$\frac{k(Q-q)}{b} = \frac{kq'}{b} = \frac{kq}{a} \quad (192)$$

where q is the net charge transferred from b to a and q' is the remaining charge on b . This can be rewritten as:

$$\frac{q}{q'} = \frac{a}{b} \quad (193)$$

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:

$$E_a = \frac{kq}{a^2} \quad (194)$$

$$E_b = \frac{kq'}{b^2} \quad (195)$$

If we take the ratio of the *field strengths* we get:

$$\frac{E_a}{E_b} = \frac{q}{q'} \frac{b^2}{a^2} = \frac{b}{a} \quad (196)$$

and conclude that the *field is much stronger on the surface of the smaller conductor*. In fact, it becomes *infinite* in the limit that $a \rightarrow 0$ 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!

3.6: 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.

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, transferring 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. The field at a sharp point can easily be strong enough to ionize air molecules in the immediate vicinity of the tip and make them conduct! The ionized air molecules recover electrons from their surroundings, which emit light as they rebind. This light (visible in the dark as a faint blue-violet glow on a thumbtack point attached to an electrostatic generator) is called the *corona*.

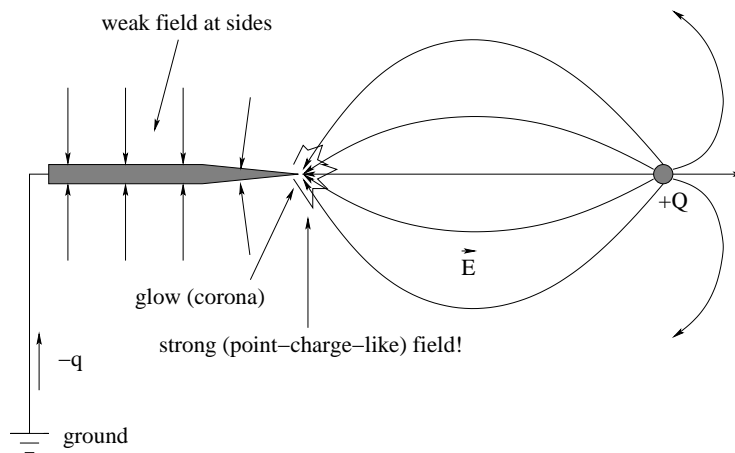


Figure 38: External charge $+Q$ induces a charge $-q$ on the sharp tip of a nearby conductor. Electric fields lines leave the tip at right angles, producing a field that looks like that of a very large *point charge* which is extremely strong very close to the tip. This in turn ionizes nearby air molecules, creating the *corona* (and spraying/repelling negatively charged ions out into the air where they are attracted to $+Q$ and eventually neutralize it).

Those molecules quickly pick up charge from the tip and are then *repelled* by it. They literally spray away from it, carrying charge and momentum and flowing towards the inducing charge. This is a process called *corona discharge* and is how lightning works. A lightning rod does not *attract* lightning (you *never* want to attract lightning) it *neutralizes it* by allowing charge to *gradually* be pulled up from the ground and sprayed onto an approaching strongly charged cloud and *slowly* neutralize it.

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 charge q at position $z = a$ on the z -axis and charge $-q$ at $z = -a$ – an *electric dipole* as studied in the first chapter. a) Write an exact expression for the electrostatic potential of the dipole at $\vec{r} = (r, \theta, \phi)$. 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, $p_z = 2qa$.

Bonus: Where is the potential of this arrangement identically zero? Right, the xy -plane. Suppose one slides an (infinite) thin *grounded* conducting plane in between the two charges. This costs no work (right?) 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, do you expect a bare charge of either sign to be attracted to or repelled by a nearby grounded conducting sheet?

Problem 3.

Now let's assume a charge $-q$ at *both* positions $z = \pm a$ on the z -axis and a charge $+2q$ at the origin. Note that this is a pair of *opposed* electric dipoles. a) Write an exact expression for the electrostatic potential of the dipole at $\vec{r} = (r, \theta, \phi)$. 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. c) What might we call this term? (Hint: Count the poles.)

Problem 4.

Find by direct integration the potential on the axis of a thin disk of charge with surface charge density σ and radius R . Then expand the result to leading order in the two limits $R \gg z$ and $z \gg R$ and interpret the potentials in both of these cases.

Problem 5.

How much work is required to assemble a uniform ball of charge with total (final) charge Q and radius R ? Hint: This is the same as the potential energy of the sphere, so use $dU = V dq$ and imagine “building” the sphere a layer of thickness dr at a time. Alternatively, compute the work directly by bringing a charge dq in from infinity against the electric field of the charge already there (and distributed as a sphere of radius r).

Problem 6.

Compute the potential difference 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*!

Problem 7.

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 $-Q$. 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 charge to come to a new equilibrium. What is: a) The charge on all three shells? b) The potential at all points in space (this is quite a bit of work, but when you’re done you’ll really have the hang of this down)?

Problem 8.

Two rings of charge Q and radius R (uniformly distributed) are located at $z = \pm R$ and have the same (z) axis. A small bead of mass m with charge q is threaded on a frictionless string along the z axis. If the bead is displaced a small distance $+z_0 \ll R$ from the origin, describe the subsequent motion of the bead in detail. (Hint: That means find $z(t)$ and the approximate period T or angular frequency ω of harmonic oscillation for the bead, in case that wasn’t clear.)

Problem 9.

Suppose you have a solid sphere with a radius R and a uniform charge density ρ . Find the potential at all points in space. Now repeat this for a *non*-uniform charge density of the form $\rho(r) = \rho_0 \frac{r}{R}$ (starting by using Gauss's Law to find the field). Note that this is *right on the edge* of being an “advanced” problem as it requires you to do an *integral* to evaluate the total charge inside a Gaussian surface. To keep it from being “just” an exercise in calculus, note the following:

The volume of a differentially thin spherical shell is its area $4\pi r'^2$ times its thickness dr' :

$$dV = 4\pi r'^2 dr'$$

The charge in this shell is therefore:

$$dQ = \rho(r') 4\pi r'^2 dr' = \frac{4\pi\rho_0}{R} r'^3 dr'$$

So integrate both sides between sensible limits to find the charge inside a Gaussian sphere of a given radius inside or outside of the sphere. You can do it! (BTW, I use r' instead of r so you can make r a limit of integration – remember how that works?)

Advanced Problem 10.

Let's try to use this to understand a little bit about nuclear fission. Suppose that the charge Q in the previous problem is distributed uniformly in an *incompressible fluid*. Now imagine that sphere splitting into two identical, smaller spheres. Find the radius R' of these two spheres. Obviously, each sphere has a charge of $Q/2$. Find the total electrostatic energy of these two spheres once they have stabilized and are separated by a large distance. 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:

$$C = \frac{|\Delta Q|}{|\Delta V|} \quad (197)$$

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:

$$1\text{F} = \frac{1 \text{ Coulomb}}{1 \text{ Volt}} \quad (198)$$

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:

$$C = \frac{\epsilon_0 A}{d} \quad (199)$$

where A is its cross sectional area and d is the separation of the plates.

- Cylindrical capacitor:

$$C = \frac{2\pi L \epsilon_0}{\ln(b/a)} \quad (200)$$

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:

$$C = 4\pi\epsilon_0 \frac{ab}{(b - a)} \quad (201)$$

where a is the outer radius of the inner conductor and b the inner radius of the outer conductor.

- Energy stored in a capacitor:

$$U = \frac{1}{2}QV = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} \quad (202)$$

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:

$$\eta_e = \frac{1}{2}\epsilon_0 E^2 \quad (203)$$

- Adding capacitors in parallel:

$$C_{\text{tot}} = C_1 + C_2 + \dots \quad (204)$$

- Adding capacitors in series:

$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + \dots \quad (205)$$

- 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:

$$\vec{E} = \frac{\vec{E}_0}{\epsilon_r} \quad (206)$$

where \vec{E}_0 is the external field, \vec{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:

$$\epsilon = \epsilon_r \epsilon_0 \quad (207)$$

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:

$$C = \epsilon_r C_0 \quad (208)$$

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:

$$E_r = \frac{k_e Q}{r^2} \quad (r > R) \quad (209)$$

$$= 0 \quad (r < R \text{ inside the conductor}) \quad (210)$$

If we integrate this to find the potential everywhere in space we get:

$$\begin{aligned} V &= - \int_{\infty}^r \frac{kQ}{r^2} dr \\ &= 0 \frac{k_e Q}{r} \quad (r \geq R) \end{aligned} \quad (211)$$

The conductor is *equipotential*, so the potential inside is the same as at its surface:

$$V = \frac{k_e Q}{R} \quad (r < R) \quad (212)$$

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 "volticance" of the conductor \mathcal{V} so that (in the case of this example):

$$V = \mathcal{V}Q \quad (213)$$

with

$$\mathcal{V} = \frac{k_e}{R} = \frac{1}{4\pi\epsilon_0 R} \quad (214)$$

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:

$$Q = CV = \left(\frac{1}{\mathcal{V}}\right) V = (4\pi\epsilon_0 R)V \quad (215)$$

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:

$$C = \frac{Q}{V} \quad (216)$$

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:

$$C = 4\pi\epsilon_0 R \quad (217)$$

In general the *SI units* of capacitance are easily remembered (as always) from the defining relation:

$$1 \text{ Farad} = \frac{1 \text{ Coulomb}}{1 \text{ Volt}}$$

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.

Example 4.1.1: Parallel Plate Capacitor

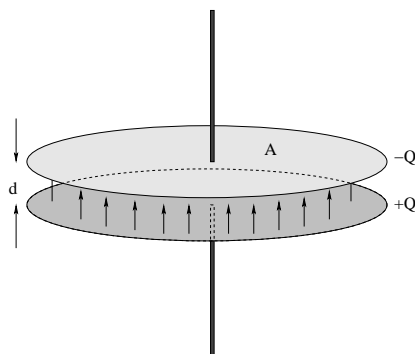


Figure 39: An “ideal” parallel plate capacitor of cross-sectional area A and plate separation d .

In figure 39 you can see the archetype for all capacitor problems. Two parallel conducting plates are arranged so that they are separated by a *small* insulating gap d (which may or may not be filled with a dielectric material, see section on dielectrics below). 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 result⁴³.

We wish to compute the capacitance, showing *all the steps*. We proceed as follows:

⁴³Why? Consider the properties of a conductor in electrostatic equilibrium, which requires perfect cancellation of the fields inside the conductors just inside the opposing surfaces...

- 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:

$$\begin{aligned} \oint_S \vec{E} \cdot \hat{n} dA &= 4\pi k_e Q_{\text{in}S} \\ |E_z| 2A &= \frac{\sigma A}{\epsilon_0} \\ E_z &= \frac{\sigma}{2\epsilon_0} = 2\pi k_e \sigma \end{aligned} \quad (218)$$

(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. Above and below both sheets, the fields produced by the upper and lower charges *cancel*, as e.g. field from the upper one points down and the field from the lower one 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:

$$E_z = 4\pi k_e \sigma = \frac{\sigma}{\epsilon_0} \quad (219)$$

directed *upwards* between the plates, as drawn, and $E_z = 0$ above and below the plates. Note well that this field is automatically *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.

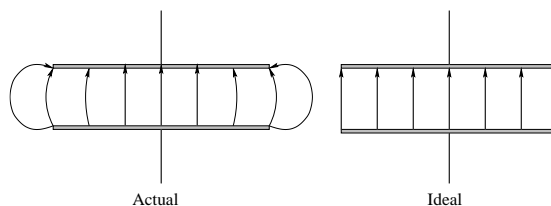


Figure 40: Fringe fields at the edge of an actual pair of parallel plates carrying opposite charge compared to the idealized field that vanishes sharply at the edge and is uniform in between the plates. Note that the field, and hence the potential difference, is almost identical in most of the volume between the plates.

At the edges of the plate, 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 a sort of dipole). This “fringing field” rapidly falls off in magnitude compared to its strength between the plates, and in this course we will *always* idealize this by asserting that the field “vanishes” at and outside of the edges of the plates and is perfectly uniform in between, even though this isn’t precisely true. This situation is portrayed in figure 40

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):

$$V = \Delta V = - \int_0^d E_z dz = -4\pi k_e \sigma d = -\frac{Qd}{\epsilon_0 A} \quad (220)$$

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:

$$C = \frac{|Q|}{|V|} = \frac{Q}{\frac{Qd}{\epsilon_0 A}} = \frac{\epsilon_0 A}{d} \quad (221)$$

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 4.1.2: 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.

As before, assume that they are charged up to $+Q$ on the inner and $-Q$ on the outer by means of 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. From Gauss’s Law it is easy to show that:

$$E_r = \frac{2k_e \lambda}{r} \quad a < r < b$$

and $E_r = 0$ otherwise (idealizing by neglecting the fringing fields that might exist at the ends of the cylinders). Then:

$$V = \Delta V = - \int_a^b E_r dr = -2k_e \lambda \ln \left(\frac{b}{a} \right) = -\frac{1}{2\pi\epsilon_0} \frac{Q}{L} \ln \left(\frac{b}{a} \right) \quad (222)$$

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:

$$C = \frac{Q}{V} = \frac{2\pi\epsilon_0 L}{\ln\left(\frac{b}{a}\right)} \quad (223)$$

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 4.1.3: Spherical Capacitor

Similarly, we can do two concentric spherical conducting shells of radius a and b , charged to $\pm Q$ on inner and outer shell respectively by our intrepid devil. From Gauss's Law:

$$E_r = \frac{k_e Q}{r^2} \quad a < r < b$$

and $E_r = 0$ otherwise, with *no* idealization or fringing fields. From this we trivially find:

$$\begin{aligned} V &= \Delta V = - \int_b^a E_r dr \\ &= k_e Q \left\{ \frac{1}{a} - \frac{1}{b} \right\} \\ &= k_e Q \left\{ \frac{b-a}{ab} \right\} \\ &= \frac{1}{4\pi\epsilon_0} Q \left\{ \frac{b-a}{ab} \right\} \end{aligned} \quad (224)$$

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:

$$C = \epsilon_0 \frac{4\pi ab}{b-a} \quad (225)$$

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 ΔQ 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 the total charge on capacitor plates is $\pm Q$, so that:

$$V = \frac{Q}{C} \quad (226)$$

is the potential difference between the plates. Then the *next* fork full of charge that he moves over, he will have to do work:

$$\Delta W = V \Delta Q \quad (227)$$

The work the *blue devil* does charging up the plates is *equal* to the change in the potential energy of the charged plates⁴⁴. We make the chunk of charge being moved differentially small, and write:

$$dU = V dQ = \frac{Q}{C} dQ \quad (228)$$

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 :

$$U = \int dU = \frac{1}{C} \int_0^{Q_0} Q dQ = \frac{1}{2} \frac{Q_0^2}{C} \quad (229)$$

We can thus easily write the total energy stored *three ways*:

$$U = \frac{1}{2} \frac{Q_0^2}{C} = \frac{1}{2} C V_0^2 = \frac{1}{2} V_0 Q_0 \quad (230)$$

(where note, we use $Q_0 = C V_0$ to go from the first to the second, then use it again to go to the third).

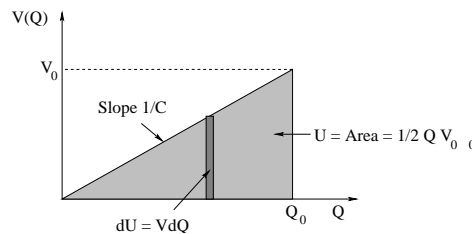


Figure 41: The energy as the area underneath the curve $V(Q) = Q/C$.

Of these, the third form is perhaps the most revealing and convenient. 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 41, $1/2 Q_0 V_0$. It's also a good time to remind you that we *did* an integral of this sort in the chapter on potential and energy, except this time we didn't distribute the charge Q in a ball, we left it in a thin layer on the surface of the capacitor plate(s) so that it is even easier (and gives us the promised factor of $1/2$ instead of $3/5$).

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, and then note that

⁴⁴Think 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.

the result can be shown to hold in the more general case of varying fields using more calculus in a later course. In this course, we will limit ourselves to *verifying* that the result is *consistent* with the energy computed for e.g. spherical or cylindrical capacitors, or with just the energy stored creating a ball of charge like the one above. This isn't quite a proof that it is general, but it certainly seems as though it makes it more likely.

Consider, then, the energy stored in a parallel plate capacitor and write it in terms of the electric field strength:

$$\begin{aligned} U &= \frac{1}{2}CV^2 = \frac{1}{2}\frac{\epsilon_0 A}{d}(Ed)^2 \\ &= \frac{1}{2}\epsilon_0 E^2(Ad) = \frac{1}{2}\epsilon_0 E^2(\text{Vol}) \end{aligned} \quad (231)$$

where Ad is the volume of the region in between the plates where the field is nonzero in our idealized picture (neglecting fringing fields). If we divide both sides of this equation by the volume, we obtain:

$$\eta_e = \frac{dU}{dV} = \frac{1}{2}\epsilon_0 E^2 \quad (232)$$

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 case we just solved, the energy of a ball of uniform charge. We write:

$$\begin{aligned} dU &= \eta_e dV = \frac{1}{2}\epsilon_0 E(r)^2 4\pi r^2 dr \\ U = \int dU &= \int \eta_e dV = \frac{1}{2}\epsilon_0 \int_0^\infty E(r)^2 4\pi r^2 dr \\ &= \frac{1}{2}(4\pi\epsilon_0) \left\{ \int_0^R \left(\frac{k_e Q}{R^3} r \right)^2 r^2 dr + \int_R^\infty \left(\frac{k_e Q}{r^2} \right)^2 r^2 dr \right\} \\ &= \frac{1}{2} \frac{1}{k_e} k_e^2 Q^2 \left\{ \int_0^R \frac{r^4}{R^6} dr + \int_R^\infty \frac{1}{r^2} dr \right\} \\ &= \frac{1}{2} k_e Q^2 \left\{ \frac{1}{5R} + \frac{1}{R} \right\} = \frac{1}{2} k_e Q^2 \frac{6}{5R} \\ &= \frac{3}{5} \frac{k_e Q^2}{R} \end{aligned} \quad (233)$$

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.

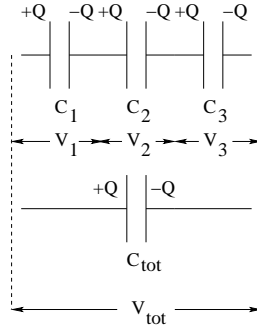


Figure 42: Find the total capacitance of a bunch of capacitors *in series*.

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 42 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_{tot} .

We need to find what C_{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 arrangement, 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 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 equilibrium 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 $V_{\text{tot}} = Q/C_{\text{tot}}$, 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:

$$V_{\text{tot}} = V_1 + V_2 + V_3 + \dots = \frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3} + \dots \quad (234)$$

$$V_{\text{tot}} = \frac{Q}{C_{\text{tot}}} \quad (235)$$

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:

$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots = \sum_i \frac{1}{C_i} \quad (236)$$

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 “volticittance” 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 common

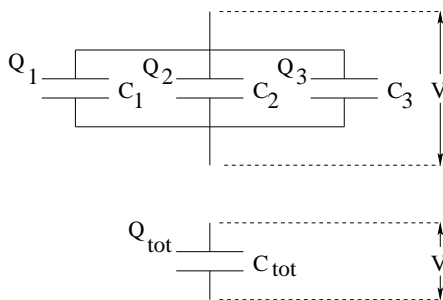


Figure 43: Find the total capacitance of a much of capacitors *in parallel*.

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 43 *each* capacitor in the top arrangement has a potential V across it. Therefore the first capacitor has a charge $Q_1 = C_1 V$, the second has a charge $Q_2 = C_2 V$, the third $Q_3 = C_3 V$. The equivalent *total* capacitance C_{tot} with the *same* voltage V across it has a charge $Q_{\text{tot}} = C_{\text{tot}} V$ 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!

$$\begin{aligned} Q_{\text{tot}} &= Q_1 + Q_2 + Q_3 + \dots \\ C_{\text{tot}} V &= C_1 V + C_2 V + C_3 V + \dots \\ C_{\text{tot}} &= C_1 + C_2 + C_3 + \dots = \sum_i C_i \end{aligned} \quad (237)$$

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 increase 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:

$$F_x = -k_s x \quad (238)$$

$$V = -\frac{1}{C} Q \quad (239)$$

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. So 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 44:

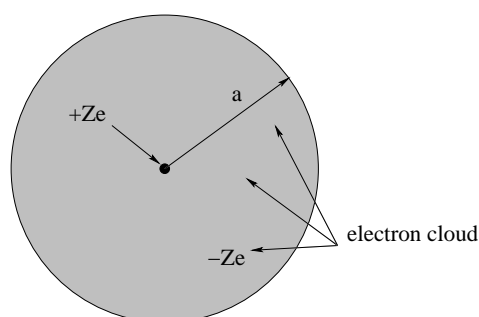


Figure 44: An “atom” consisting of a tiny massive nucleus surrounded by a *uniform* ball of negative charge modelling the “electron cloud”.

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 displaces 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:

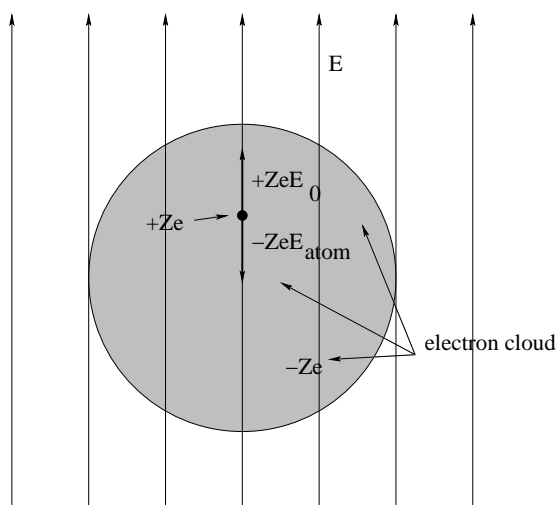


Figure 45: An “atom” polarized by an external electric field.

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):

$$E_{\text{atom}} = \frac{-k_e(Ze)z}{a^3} \quad (240)$$

(down). Thus the forces balance when:

$$+ ZeE_0 - \frac{k_e(Ze)^2 z_0}{a^3} = 0 \quad (241)$$

We can then solve for the dipole moment of the polarized atom:

$$p_z = (Ze)z_0 = \frac{a^3}{k_e} E_0 = 4\pi\epsilon_0 a^3 E_0 \quad (242)$$

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:

$$P = \frac{p_z}{8a^3} = \frac{\pi}{2}\epsilon_0 E_0 \quad (243)$$

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 guesstimate 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:

$$\vec{P} = \chi_e \epsilon_0 \vec{E} \quad (244)$$

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!)

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!

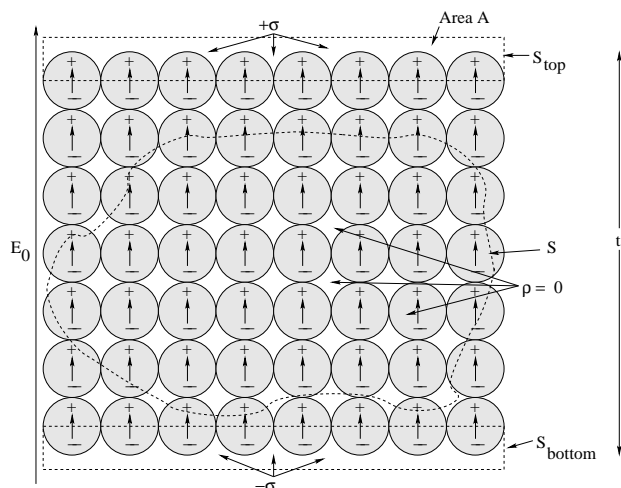


Figure 46: A lattice of atoms polarized by an external electric field.

In figure 46, 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 us 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*⁴⁵.

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_{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_{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**

⁴⁵If it contained an integer number of whole atoms, it would be exactly zero. If the surface cuts through atoms 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.

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.

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:

$$p_{\text{system}} = Q_{\text{surface}}t = (\sigma_b A)t = PV = P(At) \quad (245)$$

(all in the direction of the field) or clearly:

$$\sigma_b = P \quad (246)$$

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 accommodate 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 fields⁴⁶.

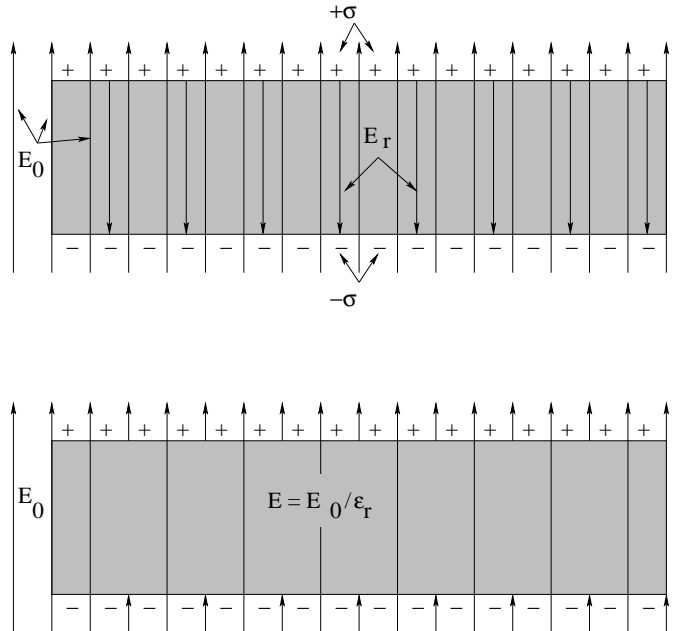


Figure 47: The polarized material generates a *reaction field* E_r that *opposes* the applied field and partially cancels it, making the total field in the material smaller. A dielectric material thus *reduces* the applied electric field inside the material.

⁴⁶Truly 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.

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 47.

Applying Gauss's Law to the induced surface charge layers in this simple rectangular geometry, we expect:

$$E_r = \frac{\sigma_b}{\epsilon_0} \quad (247)$$

The total field is then:

$$E = E_0 - \frac{\sigma_b}{\epsilon_0} = E_0 - \frac{P}{\epsilon_0} = E_0 - \chi_e E \quad (248)$$

We can rearrange this into:

$$E(1 + \chi_e) = E_0 \quad (249)$$

and solve for E , the field inside the material, in terms of E_0 , the applied external field:

$$E = \frac{E_0}{1 + \chi_e} = \frac{E_0}{\epsilon_r} \quad (250)$$

where we have introduced the *relative permittivity*

$$\epsilon_r = (1 + \chi_e) \quad (251)$$

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 κ to represent $1 + \chi_e$ and call it the *dielectric constant* for the material, as in:

$$\kappa = (1 + \chi_e) = \epsilon_r \quad (252)$$

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{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{E} by $\vec{P} = \chi\epsilon_0\vec{E}$. The net field itself is related to the applied field by $\vec{E} = \vec{E}_0/\epsilon_r$ 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***:

$$\epsilon = \epsilon_r\epsilon_0 \quad (253)$$

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**:

$$\vec{D} = \epsilon \vec{E} \quad (254)$$

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:

$$\oint_S \vec{D} \cdot \hat{n} dA = \int_{V/S} \rho_f dV \quad (255)$$

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{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:

$$\vec{D} = \frac{1}{4\pi} \frac{Q}{r^2} \hat{r} \quad (256)$$

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 and 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.

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:

$$E_{\text{material}} = \frac{E_0}{\epsilon_r} \quad (257)$$

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 strengths.

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:

- They mechanically separate the plates.
- They increase the capacitance.
- 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:

$$E_0 = 4\pi\epsilon_0\sigma_f = \frac{\sigma_f}{\epsilon_0} \quad (258)$$

However, now there *is* a dielectric in that space. The field is modified to become:

$$E = \frac{E_0}{\epsilon_r} = \frac{\sigma_f}{\epsilon_r\epsilon_0} = \frac{\sigma_f}{\epsilon} \quad (259)$$

Next, we compute as usual the potential difference:

$$V = - \int_d^0 \frac{Q_f}{A\epsilon_r\epsilon_0} dz = \frac{Q_f d}{A\epsilon_r\epsilon_0} \quad (260)$$

and the capacitance:

$$C = \frac{Q_f}{V} = \epsilon_r \frac{\epsilon_0 A}{d} = \epsilon_r C_0 \quad (261)$$

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 preventing 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

$$\epsilon_0 = 8.85 \times 10^{-12} \approx 10^{-11} \frac{\text{farads}}{\text{meter}} \quad (262)$$

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.

In figure 48 we can write the field in the dielectric in two ways:

$$E = \frac{E_0}{\epsilon_r} = E_0 - E_r \quad (263)$$

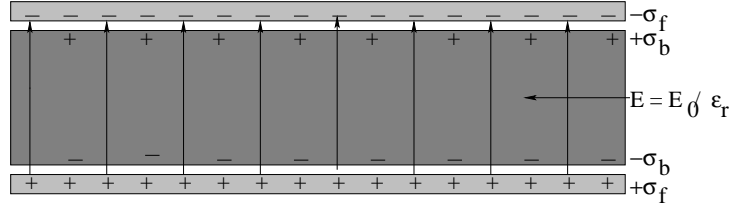


Figure 48: Bound and free charge in a capacitor filled with a dielectric.

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:

$$\frac{4\pi k_e \sigma_f}{\epsilon_r} = 4\pi k_e \sigma_f - 4\pi k_e \sigma_b \quad (264)$$

If we cancel out the common factor of $4\pi k_e = 1/\epsilon_0$, we get:

$$\frac{\sigma_f}{\epsilon_r} = \sigma_f - \sigma_b \quad (265)$$

or

$$\begin{aligned} \sigma_b &= \left(1 - \frac{1}{\epsilon_r}\right) \sigma_f \\ &= \left(\frac{\epsilon_r - 1}{\epsilon_r}\right) \sigma_f \\ &= \left(\frac{-\chi}{1 + \chi}\right) \sigma_f \end{aligned} \quad (266)$$

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).

$$\begin{aligned} \sigma_b &= \epsilon_0 \chi_e E \\ &= \epsilon_0 \chi_e \frac{E_0}{\epsilon_r} \\ &= \epsilon_0 \chi_e \frac{\sigma_f}{\epsilon_0 \epsilon_r} \\ &= \sigma_f \frac{\chi_e}{1 + \chi_e} \end{aligned} \quad (267)$$

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 49:

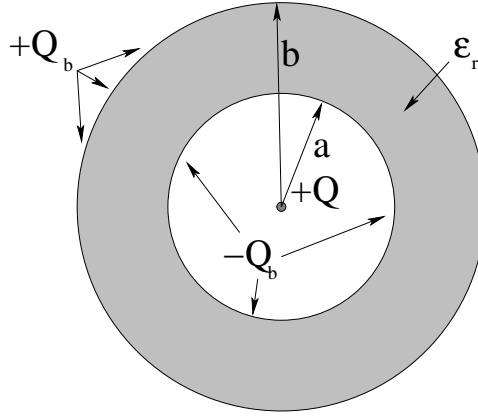


Figure 49: A bare charge $+Q$ surrounded by a dielectric shell with relative permittivity ϵ_r .

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:

$$\sigma_b = -\chi_e \frac{Q}{(4\pi a^2)\epsilon_r} = -\frac{Q(\epsilon_r - 1)}{(4\pi a^2)\epsilon_r} \quad (268)$$

$$\sigma_b = +\chi_e \frac{Q}{(4\pi b^2)\epsilon_r} = +\frac{Q(\epsilon_r - 1)}{(4\pi b^2)\epsilon_r} \quad (269)$$

Note well that the *total* bound charge on *either* surface has magnitude:

$$Q_b = Q \frac{\epsilon_r - 1}{\epsilon_r} \quad (270)$$

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{n} is a normal unit vector perpendicular to a dielectric surface, where the polarization density is e.g. $\vec{P} = \epsilon_0 \chi_e \vec{E}$ for \vec{E} just **inside** the material. Then σ_b , which is a **scalar**, is given by:

$$\sigma_b = \vec{P} \cdot \hat{n}$$

Our treatment above was valid for the special case that $\hat{n} \parallel \vec{P}$, 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:

$$\vec{\nabla} \cdot \vec{P} = -\rho_b$$

or equivalently:

$$\oint_S \vec{P} \cdot \hat{n} dA = - \int_{V/S} \rho_b dV$$

(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:

$$\vec{E} = \vec{E}_0 - \frac{\vec{P}}{\epsilon_0}$$

(see above) and if we take the divergence of both sides:

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho_{\text{tot}}}{\epsilon_0} = \vec{\nabla} \cdot \vec{E}_0 - \frac{\vec{\nabla} \cdot \vec{P}}{\epsilon_0} = \frac{\rho_f + \rho_b}{\epsilon_0}$$

where we used $\vec{\nabla} \cdot \vec{E}_0 = \rho_f/\epsilon_0$ 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:

$$U = \frac{1}{2} CV^2 = \frac{1}{2} \frac{\epsilon_r \epsilon_0 A}{d} (Ed)^2 \quad (271)$$

where E is still the field between the plates, in this case the field inside the dielectric. Hence

$$\eta_e = \frac{dU}{dV} = \frac{1}{2} \epsilon E^2 \quad (272)$$

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 parallel plate capacitor with cross-sectional area A and plate separation d ;
- b) A cylindrical capacitor with inner conductor radius a , outer conductor radius b , and length L (where $L \gg b - a$);
- c) A spherical capacitor with inner conductor radius a and outer conductor radius b .

Show in the latter two cases that the capacitance is approximately:

$$C \approx \frac{\epsilon_0 A}{d}$$

where A is the area of the cylinder/sphere and $d = b - a \ll a$ ("small" separation). You will need to use the power series expansion $\ln(1 + x) \approx x + \mathcal{O}(x^2) \dots$ to first order to do the cylinder.

Problem 3.

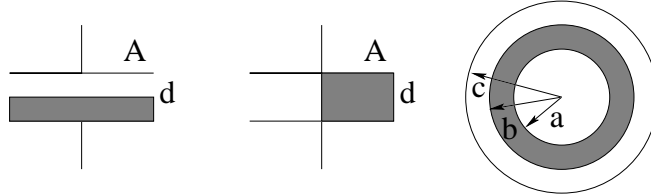
Prove that the energy stored on the capacitor can be written as *either* side of:

$$U = \frac{1}{2}QV = \int_{\mathcal{V}} \frac{1}{2}\epsilon_0 E^2 dV$$

for all three geometries (where the 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:

$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots \quad (\text{series})$$

and

$$C_{\text{tot}} = C_1 + C_2 + C_3 + \dots \quad (\text{parallel})$$

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 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

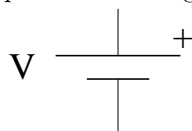
$$F_x = -\frac{dU}{dx}$$

where U is the *total* potential energy of the voltage source and capacitor, find the force on the partially inserted plate. 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:

$$I = \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt} \quad (273)$$

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:

$$\Delta V = IR \quad (274)$$

which can be modelled from:

$$R = \frac{\rho L}{A} = \frac{L}{\sigma A} \quad (275)$$

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:

$$\vec{J} = \frac{\Delta Q}{A\Delta t} \hat{n} = \sigma \vec{E} \quad (276)$$

where \vec{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:

$$P = VI = \frac{V^2}{R} = I^2 R \quad (277)$$

where the first form is the easiest to understand.

- Adding resistors in series:

$$R_{\text{tot}} = R_1 + R_2 + \dots \quad (278)$$

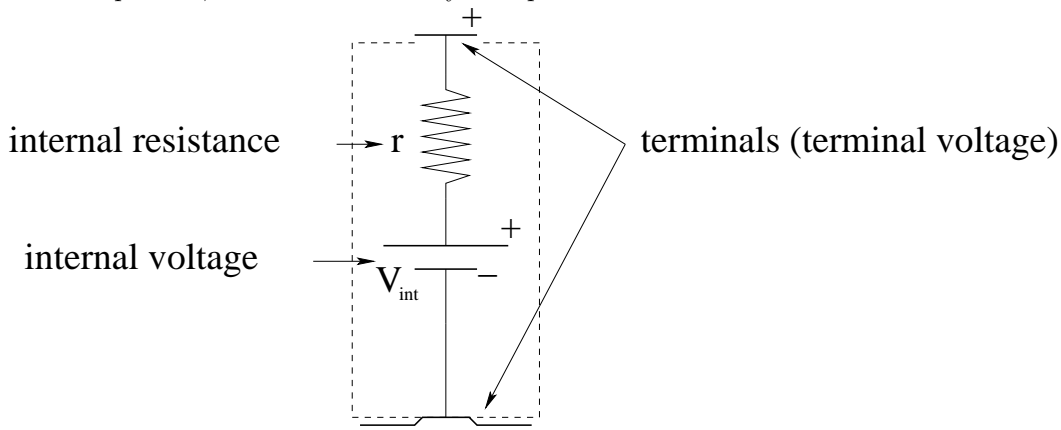
- Adding resistors in parallel:

$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots \quad (279)$$

- Kirchhoff's Rules:

- Loop Rule:** The sum of the voltage changes around a circuit *loop* must be zero (conservation of energy).
- 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:

$$V_{\text{terminal}} = V_{\text{int}} - Ir \quad (280)$$

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:

$$I_{\text{max}} = \frac{V_{\text{int}}}{r} \quad (281)$$

and:

$$P_{\text{max}} = V_{\text{int}} I_{\text{max}} = \frac{V_{\text{int}}^2}{r} \quad (282)$$

- 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**:

$$V_C(t) = V_0 e^{-t/RC} \quad (283)$$

or:

$$Q(t) = Q_0 e^{-t/RC} \quad (284)$$

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:

$$V_C(t) = V_0 \left(1 - e^{-t/RC}\right) \quad (285)$$

or:

$$Q(t) = Q_0 \left(1 - e^{-t/RC}\right) \quad (286)$$

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*⁴⁷.

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 anode⁴⁸ :



while at the cathode:



or overall:



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 *favoured*.

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

⁴⁷Technically, 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".

⁴⁸Wikipedia: 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.

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).

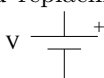
A second advantage of this particular battery is that it is *rechargeable*. If one simply places a voltage across the cell that exceeds its terminal voltage, charge flows *the other way*, reversing the reaction and turning lead sulphate back into lead or lead oxide. By careful design, one can charge 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 battery⁴⁹, 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 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.

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 , where V 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.

⁴⁹<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.

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.

5.1.1: 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, rechargeable 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 intermittent 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, rechargeable 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 rechargeable/secondary battery technology (which is very much a moving target, where the costs per unit of energy stored by a rechargeable battery have decreased by some 60 to 80% over the last ten or fifteen years) by visiting:

Wikipedia: [http://www.wikipedia.org/wiki/Rechargeable Battery](http://www.wikipedia.org/wiki/Rechargeable_Battery)

To summarize: at this point (with this paragraph being written in 2017) large capacity, high density, long lifetime rechargeable 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 manufacturing 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 actually 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 day 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.

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 electrostatic 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.

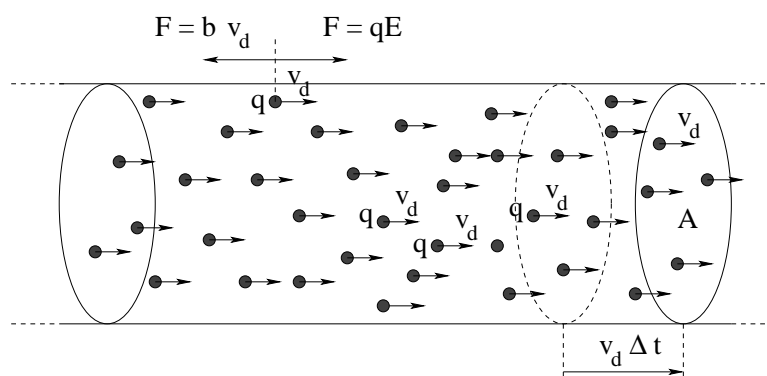


Figure 50: The simple linear model for conduction in a resistive lattice.

In figure (50) 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:

$$v_d = \frac{qE}{b} \quad (287)$$

where we introduce the **drift velocity** v_d , defined to be the average “terminal velocity” of charges in the conductor⁵⁰. 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 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.

⁵⁰We 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).

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:

$$n = \frac{\# \text{ of charge carriers}}{\text{unit volume}} \quad (288)$$

the number of (free!) charge carriers q per unit volume. We can then turn this into the *free charge density*:

$$\rho_{\text{free}} = nq \quad (289)$$

Using these quantities, we see that:

$$\Delta Q = nqv_d A \Delta t = \rho_{\text{free}} v_d A \Delta t \quad (290)$$

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:

$$I = \frac{\Delta Q}{\Delta t} \approx \frac{dQ}{dt} = nqv_d A = \rho_{\text{free}} v_d A \quad (291)$$

In passing we note that the SI units of current are *Amperes* (or Amps for short) where

$$1 \text{ Ampere} = \frac{1 \text{ Coulomb}}{1 \text{ Second}} \quad (292)$$

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.

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) steady-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 $A' = A/\cos(\theta)$ 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:

$$I = nqAv_d = nq\frac{A}{\cos(\theta)}\cos(\theta)v_d = nqA'v_d\cos(\theta) \quad (293)$$

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{n} , the *directed normal unit vector* to the surface A or A' :

$$I = nqA\vec{v}_d \cdot \hat{n} = nqA'\vec{v}_d \cdot \hat{n}' \quad (294)$$

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{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{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{v}_d . For each of these, the differential current through the chunk is:

$$dI = nq\vec{v}_d \cdot \hat{n}dA \quad (295)$$

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:

$$I_C = \int_{S/C} nq\vec{v}_d \cdot \hat{n}dA = \int_{S/C} \vec{J} \cdot \hat{n}dA \quad (296)$$

where S/C is read “through the surface S bounded by the closed curve C and where:

$$\vec{J} = nq\vec{v}_d = \rho_{\text{free}}\vec{v}_d \quad (297)$$

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:

$$Q_{\text{in } S} = \int_{V/S} \rho dV \quad (298)$$

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{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!):

$$I_{\text{out}} = \oint_S \vec{J} \cdot \hat{n} dA = -\frac{d}{dt} \int_{V/S} \rho dV = -\frac{dQ_{\text{in } S}}{dt} \quad (299)$$

which we rearrange as:

$$\oint_S \vec{J} \cdot \hat{n} dA + \frac{d}{dt} \int_{V/S} \rho dV = 0 \quad (300)$$

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 outside⁵¹.

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:

$$\vec{\nabla} \cdot \vec{J} + \frac{\partial \rho_e}{\partial t} = 0. \quad (301)$$

This ends up being much more convenient for doing the math associated with solving serious electrodynamics problems. It 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.

⁵¹There 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.

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 the following picture: In this figure, we

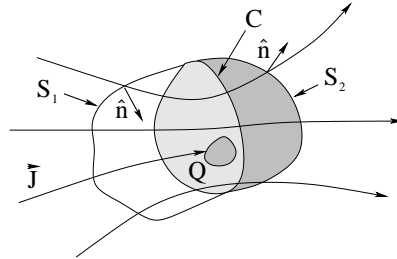


Figure 51: Charge Conservation.

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” (in analogy with a worked problem) placed face to face, through which charge can flow as illustrated in figure 51.

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 through S_2 is the rate at which charge builds up inside V :

$$\int_{S_1} \vec{J} \cdot \hat{n} dA - \int_{S_2} \vec{J} \cdot \hat{n} dA = I_{\text{into } V \text{ thru } S_1} - I_{\text{out of } V \text{ thru } S_2} = \frac{d}{dt} \int_{V/S} \rho_e dV = \frac{dQ_{\text{in } S/V}}{dt} \quad (302)$$

Note that all we really did to get this result is split the integral over the closed surface S 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 point it out and suggest that you give it a try.

The Drude Model

The earliest forms of Wikipedia: <http://www.wikipedia.org/wiki/pinball> 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 52. 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 box a ball would end up in, or try to use skill in the way the ball was dropped to determine the outcome, but because the (essentially classical) motion was effectively chaotic, it

was effectively impossible to drop a ball in such a way that the final outcome could be predicted or controlled.

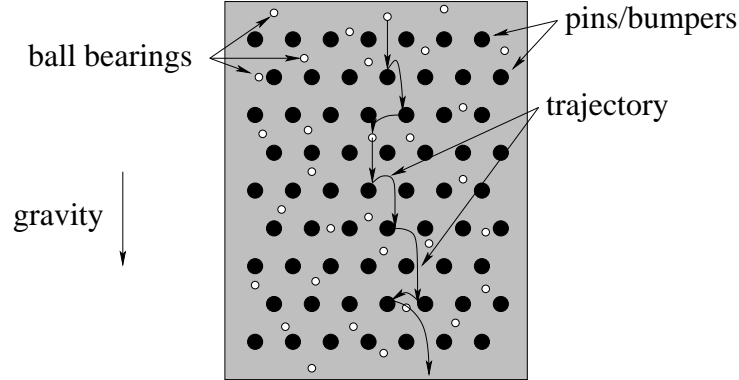


Figure 52: An early pinball machine. Balls (typically small ball bearings) dropped in at the top fall into an array of “pins” that function as bumpers but are “stopped” by the pins after falling a short time τ so they only build up a finite downward average speed.

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.

Physicists and mathematicians got involved in the game at a very early point – for example the Wikipedia: [http://www.wikipedia.org/wiki/Bean Machine](http://www.wikipedia.org/wiki/Bean_Machine) 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 above, the balls fall between pins only due to the force of gravity, and the pins themselves effectively stop their downward motion so they have to build up speed again after each collision. In a lattice of atoms, the atoms are at a *finite temperature* and the electrons are in *thermal equilibrium* (more or less) with the lattice. This means that the average thermal kinetic energy of the electrons is much, much larger than the energy they might gain from the field between collisions!

To put it another way, if it takes an electron a time τ_E to “fall” (say) some average distance between atoms/collisions and a time τ_{therm} for the atom to travel that same distance due to their average speed due to their temperature:

$$\tau_E \gg \tau_{\text{therm}} \quad (303)$$

During the shorter time τ_{therm} , the component of the velocity of an electron in the direction of the force due to the 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.

The resulting “pinball model”, pictured in figure 53, is called the **Drude model**. With no voltage/field, this is basically a horizontal “active” pinball table filled with pinballs (charges) with bumpers (atoms) that are firing/vibrating at a rapid rate so that charges constantly bounce between

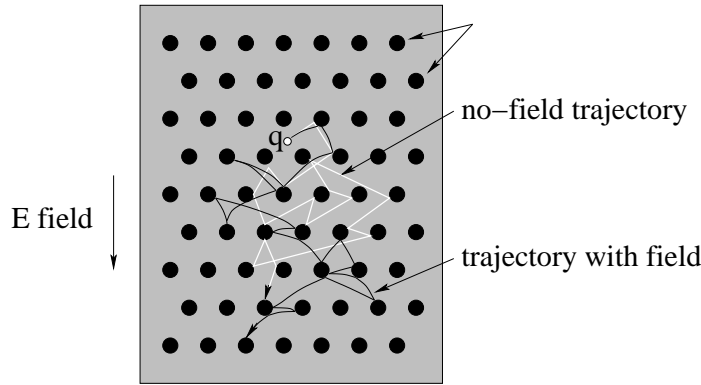


Figure 53: The Drude model: With no field, the charges q bounce very rapidly between atomic “bumpers” that maintain a roughly thermal distribution of charge speeds. On average, these collisions form a “random walk” with no direction and zero average displacement. With a field, in the very short time 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.

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 furiously vibrating pinball table* up through a small angle relative to gravity. In a pinball table, this would cause the balls to gradually drift lower in the direction of the net gravitational field. In a conductor, there is a small net acceleration in the direction determined by the electrostatic force on the charges during the *short* time between bounces. The charges all “bounce” tens to thousands of times between the atoms in the time that it would take a charge to *freely fall* from one atom to the next due only to the applied field, 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 constantly take pinballs off of the lower end of the table and elevate them back up to the top to start over so that charge doesn’t accumulate at the bottom and generate a field that stops the process. This constant lifting is an excellent model of a *battery*!

With this insight, let’s generate the algebraic description of the Drude model. We start by estimating the mean speed of the 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:

$$K = \frac{3}{2}k_bT = \frac{1}{2}m(\langle v_x \rangle^2 + \langle v_y \rangle^2 + \langle v_z \rangle^2) = \frac{1}{2}m \langle v \rangle^2 \quad (304)$$

Then we expect $\langle v \rangle$ to be:

$$\langle v \rangle = \sqrt{\frac{3k_bT}{m}} \quad (305)$$

This speed is easy enough to estimate at (say) 300 degrees kelvin and for electrons from $k_b = 1.38 \times 10^{-23}$ J/K and $m_e = 9.1 \times 10^{-31}$ as $\langle v \rangle \approx 10^5$ m/sec.

Note that this is *very, very fast*, and at that is **not fast enough** because the electrons are *not* a classical gas of non-interacting particles, they are quantum mechanical fermions whose effective

“speed” is determined by something called the fermi energy that is only weakly dependent on the temperature and leads to average speeds roughly an order of magnitude larger:

$$\langle v \rangle \sim 10^5 \text{ m/sec} \quad (306)$$

Note that this is an appreciable fraction of the speed of light and two orders of magnitude larger than escape velocity from the surface of the earth!

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 estimation) that $d = \langle v \rangle \tau_{\text{therm}} \approx 10^{-10}$ meters or one angstrom, a typical 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:

$$\tau_{\text{therm}} = \frac{d}{\langle v \rangle} \sim 10^{-16} \text{ seconds} \quad (307)$$

In passing, let’s compare this to the time τ_E it might 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 above. Since we want a quantitative estimate, 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:

$$E = \frac{\Delta V}{\ell} = \frac{100 \text{ volts}}{0.01 \text{ meters}} \approx 10^4 \text{ volts/meter.} \quad (308)$$

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. From E we can easily estimate the expected acceleration a :

$$a = qE/m = 1.6 \times 10^{-19} * 10^4 / 9.1 \times 10^{-31} = 1.76 \times 10^{15} \text{ m/sec}^2 \quad (309)$$

To estimate τ_E , we recall that (neglecting factors of order unity):

$$\tau_E = \sqrt{\frac{2md}{qE}} = \sqrt{\frac{2d}{a}} \sim 10^{-13} \text{ seconds} \quad (310)$$

which is orders of magnitude greater than τ_{therm} as expected even for very strong fields.

During the short time τ_{therm} 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 momentum gained during τ_{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 into the entire material while remaining in thermal equilibrium with it.

We are finally ready to build the model. We expect the equation of motion of our “thermalized pinball” **between collisions** to be something like:

$$\Delta \langle \vec{p} \rangle = m \Delta v = \vec{F} \tau_{\text{therm}} = q \vec{E} \tau_{\text{therm}} \quad (311)$$

so that the average acceleration of a particle between collisions will be:

$$\langle \vec{a}_c \rangle = \frac{\Delta \langle \vec{p} \rangle}{m \tau_{\text{therm}}} = \frac{q \vec{E}}{m} \quad (312)$$

Since this acceleration only applies for the average time τ_{therm} before the charge is redirected in a random direction with the unchanged thermal distribution of speeds, the average velocity during this time is:

$$\langle \vec{v} \rangle = \frac{1}{2} \langle \vec{a}_c \rangle \tau_{\text{therm}} = \frac{1}{2} \frac{q \vec{E}}{m} \tau_{\text{therm}} = \vec{v}_d \quad (313)$$

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 trying to compute τ using $a = qE/m$ starting *from rest*, we use the average time τ between collisions driven by the strong interaction of the charged particles with the surrounding material. This seems like a small change, but it is a very important one!

In this “active pinball” Drude model, then, we expect that:

$$\vec{J} = nq\vec{v}_d = \frac{nq^2\tau_{\text{therm}}}{m}\vec{E} = \sigma\vec{E} \quad (314)$$

which scales **linearly with the electric field strength**.

If we used time τ_E from the naive *passive* pinball model in exactly the same argument, we obtain an average speed of $\langle v \rangle = \frac{1}{2}\vec{a}\tau_E = \frac{1}{2}\frac{q\vec{E}}{m}\tau_E = \sqrt{\frac{dqE}{2m}} = v_d$ or $J = nq\langle v \rangle = nqA\sqrt{\frac{dqE}{2m}}$ (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! The passive pinball model fails to empirically scale correctly in several ways.

If we compare this to the equilibrium condition $qE = bv_d$ we see that the “linear drag coefficient” is given by:

$$b = \frac{m}{\tau_{\text{therm}}} \quad (315)$$

or:

$$\vec{F}_d = \frac{m\vec{v}_d}{\tau_{\text{therm}}} = \frac{\Delta\vec{p}}{\Delta t} \quad (316)$$

This is conceptually perhaps the easiest way to see what’s going on. The “drag force” equals the **average momentum change per unit time** of the free charges as they move through the lattice of atoms. We’ve simply done a very simple quantitative estimate of that momentum change in terms of our active pinball model. Sadly, this also reveals that we haven’t actually made much useful progress as we’ve exchanged *one* unknown, dimensioned number characteristic of the material (b) for another (τ_{therm} and/or m).

Ohm’s Law

At last we are set to establish the connection between the (empirical!) Ohm’s Law and the Drude conduction model. We just showed above that the current density \vec{J} is proportional to the applied electric field E . In so doing, we wrapped up all of the complexity – all the unknown stuff about a conductor, including, b , n , q , m , τ_{therm} – into a single parameter called the **conductivity**:

$$\sigma = \frac{nq^2\tau_{\text{therm}}}{m} = \frac{nq^2}{b} = \frac{q\rho_{\text{free}}}{b} = \frac{1}{\rho} \quad (317)$$

where we have also defined its reciprocal ρ , called the **resistivity**⁵²

The resistivity and/or conductivity⁵³ is a characteristic of the material of the conductor in question, and as the Drude model suggests, depends on many things. Its most important dependence

⁵²Unfortunately the common symbol for resistivity is ρ , which you can easily confuse with the charge density which *also* is given the symbol ρ in most physics books. I’ve tried pretty hard to label the charge density ρ_e , read “the density of ELECTRIC charge” or $\rho_{\text{free}} = nq$, read “the density of the FREE electric charge”, and will continue to do so wherever there is any chance of confusing the two. Sadly, things are no better if we use the conductivity σ , easily confused with the surface charge density σ_e . As you gain experience, the meaning of any symbol intended/needed will generally be clear from context without the need for hints such as this.

⁵³Wikipedia: 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.

is probably upon temperature – we expect thermal speed to increase with temperature, which in turn should decrease the time between collisions and hence decrease the average speed accumulated during this time. But *how* should it depend on temperature?

Recall that in the Drude model, $\langle v \rangle = \sqrt{3kT/m}$, hence $\tau_{\text{therm}} = d/\langle v \rangle \propto 1/\sqrt{T}$. From this we expect the conductivity to vary like:

$$\sigma = \frac{nq^2\tau_{\text{therm}}}{m} \propto 1/\sqrt{T} \quad \text{or} \quad \rho \propto \sqrt{T} \quad (318)$$

We thus expect from the Drude model that the resistance of metals should increase with temperature like the square root of the temperature (and incidentally, this alters the scaling of conductivity with the mass of the charge carriers as well).

This is incorrect – the Drude model, recall, is classical but in real materials conductivity is a quantum phenomenon. In fact, the resistivities of most materials increase approximately *linearly* with temperature, at least in the range of temperatures near room temperature. Close to room temperature the resistivity is described accurately enough by the (linearized) Taylor series expansion:

$$\rho(T) = \rho_0 \{1 + \alpha(T - T_0)\} \quad (319)$$

where 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 allow resistivity to be accurately computed across a moderate, relevant, range of temperatures by means of two tabulated quantities.

However, it is too complicated for most of our purposes here⁵⁴. In this introductory *classical* textbook we we will generally assume that $\alpha \approx 0$ so that $\rho = \rho_0$ for any given material and concentrate instead below on the simple scaling of resistance with length and area of the resistor. 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.

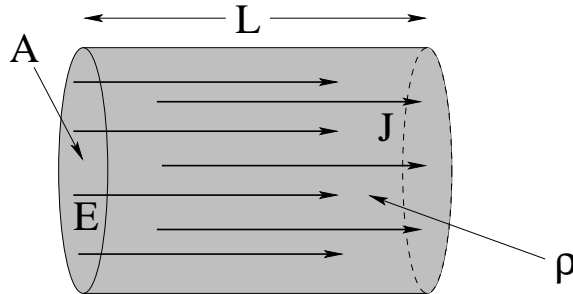


Figure 54: A simple resistor with resistivity ρ , length L , and cross sectional area A .

Now, consider an archetypical “resistor”: a uniform conductor with resistivity ρ , length L , and cross-sectional area A (where the ends are at right angles to the sides), as pictured in figure 54. We can rearrange the current density equation as:

$$\vec{E} = \rho \vec{J} \quad (320)$$

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:

$$\int \vec{E} \cdot \hat{n} dA = EA = \rho \int \vec{J} \cdot \hat{n} dA = \rho I \text{ through the resistor} \quad (321)$$

⁵⁴Unless, 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.

which we can rearrange as:

$$E = I \frac{\rho}{A} \quad (322)$$

If we integrate both sides a second time in the direction $d\vec{l}$ from one end of the conductor to the other in the direction of the current, we get the potential difference:

$$V_R = - \int \vec{E} \cdot d\vec{l} = -EL = -I \frac{\rho L}{A} = -IR \quad (323)$$

$|V_R|$ 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 the conducting material in this particular geometry:

$$R = \rho \frac{L}{A} \quad (324)$$

so that in terms of it:

$$V \text{ (or } V_R \text{)} = IR \quad (325)$$

This equation is known as *Ohm's Law* and we will use it extensively in the weeks to come.

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,

$$1 \text{ Ohm} = \frac{\text{Joule} - \text{Second}}{\text{Coulomb}^2} \quad (326)$$

Note well that the SI units of capacitance, farads, were 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 $R = \rho L/A$ 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 analogous 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 work will be complete.

For the moment, though, let's figure out how to add resistances and then study an actual dynamical problem: the *RC* circuit.

5.3: Resistances in Series and Parallel

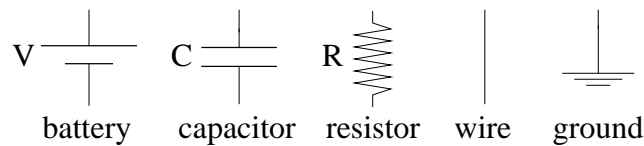


Figure 55: Symbols for batteries, capacitors, resistances, wires, and ground.

Before proceeding 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 56.

Series

Suppose we apply a fixed voltage V_{ab} across the contacts in the upper (a) diagram. This produces some current I_{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

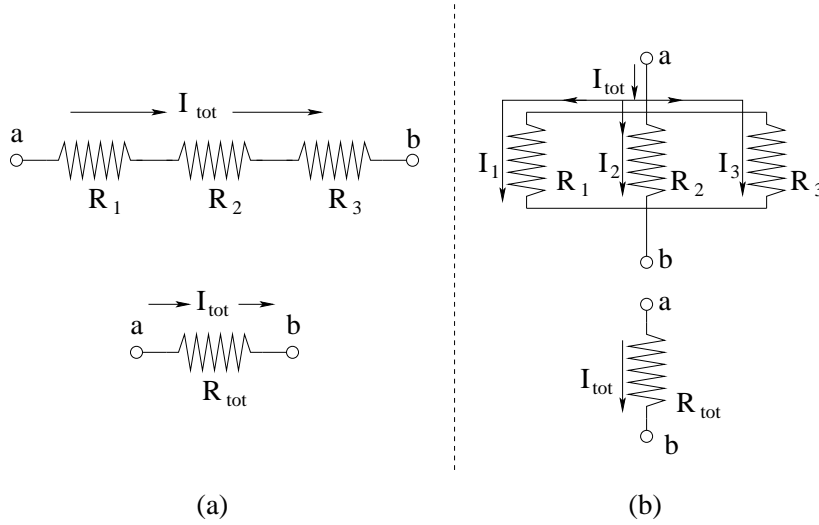


Figure 56: Three resistors R_1, R_2, R_3 arranged in *series* (left, (a)) and *parallel* (right, (b)), along with the equivalent/total resistances of each one portrayed below. In both cases the total resistance is “equivalent” when applying a voltage V_{ab} across the a and b contacts produces the *same total current* I_{tot} in the top and bottom figure.

current:

$$V_1 = I_{\text{tot}} R_1 \quad (327)$$

$$V_2 = I_{\text{tot}} R_2 \quad (328)$$

$$V_3 = I_{\text{tot}} R_3 \quad (329)$$

Obviously the total voltage V_{ab} is given by:

$$V_{ab} = V_1 + V_2 + V_3 = I_{\text{tot}}(R_1 + R_2 + R_3) \quad (330)$$

If we look at the lower (a) diagram, Ohm’s Law yields:

$$V_{ab} = I_{\text{tot}} R_{\text{tot}} \quad (331)$$

Equating and cancelling the common I_{tot} , we get:

$$R_{\text{tot}} = R_1 + R_2 + R_3 \quad (332)$$

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:

$$V_{ab} = \sum_{i=1}^N I_{\text{tot}} R_i = I_{\text{tot}} \sum_{i=1}^N R_i = I_{\text{tot}} R_{\text{tot}} \quad (333)$$

so that *in general* the rule for the addition of N resistors in series is:

$$R_{\text{tot}} = R_1 + R_2 + \dots + R_N = \sum_{i=1}^N R_i \quad (334)$$

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:

$$I_1 = \frac{V_{ab}}{R_1} \quad (335)$$

$$I_2 = \frac{V_{ab}}{R_2} \quad (336)$$

$$I_3 = \frac{V_{ab}}{R_3} \quad (337)$$

Now, consider the total current I_{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:

$$I_{\text{tot}} = I_1 + I_2 + I_3 = \frac{V_{ab}}{R_1} + \frac{V_{ab}}{R_2} + \frac{V_{ab}}{R_3} = V_{ab} \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right) \quad (338)$$

As before in the lower (b) figure we have:

$$I_{\text{tot}} = \frac{V_{ab}}{R_{\text{tot}}} \quad (339)$$

and when we equate these two forms and cancel the common V_{ab} we get:

$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \quad (340)$$

There is nothing special about three resistors, and once again we can easily generalize this argument to N resistors as:

$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_N} = \sum_{i=1}^N \frac{1}{R_i} \quad (341)$$

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 several 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 series 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 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 Rules⁵⁵.

⁵⁵Wikipedia: [http://www.wikipedia.org/wiki/Kirchhoff's Circuit Laws](http://www.wikipedia.org/wiki/Kirchhoff's_Circuit_Laws).

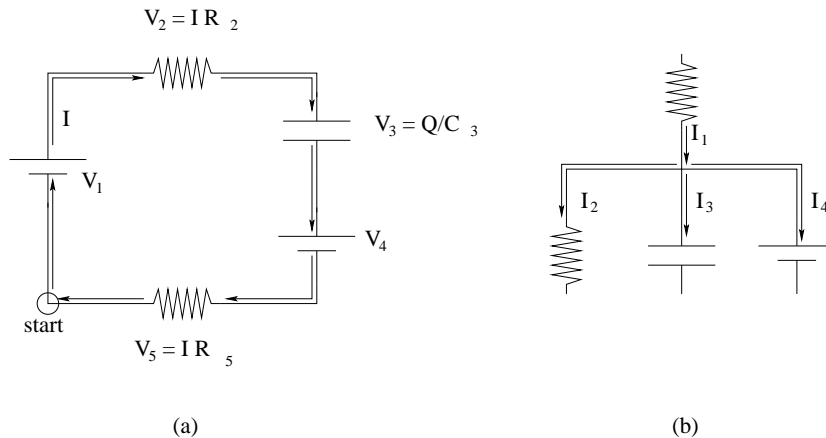


Figure 57: (a) A single “generic” circuit loop; (b) A single “generic” circuit junction.

Kirchhoff’s Loop Rule

Consider the generic *circuit loop* in figure 57 (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 (arbitrary) point “start”. As it goes across each potential V_1, V_2, \dots 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:

$$\Delta U = qV_1 + qV_2 + qV_3 + qV_4 + qV_5 = q \sum_i V_i \quad (342)$$

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:

$$\sum_i^{\text{loop}} V_i = 0 \quad (343)$$

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.

Kirchhoff’s Junction Rule

Consider the generic *circuit junction* in figure 57 (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:

$$I_{1,\text{in}} - I_{2,\text{out}} - I_{3,\text{out}} - I_{4,\text{out}} = 0 \quad (344)$$

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:

$$I_{1,\text{in}} = I_{2,\text{out}} + I_{3,\text{out}} + I_{4,\text{out}} \quad (345)$$

Note that these two equations are the same.

Let us generalize the first form to:

$$\sum_i^{\text{junction}} I_i = 0 \quad (346)$$

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:

$$\sum_i^{\text{loop}} V_i = 0 \quad \text{Loop Rule} \quad (347)$$

$$\sum_i^{\text{junction}} I_i = 0 \quad \text{Junction Rule} \quad (348)$$

Example 5.4.1: The Internal Resistance of a Battery

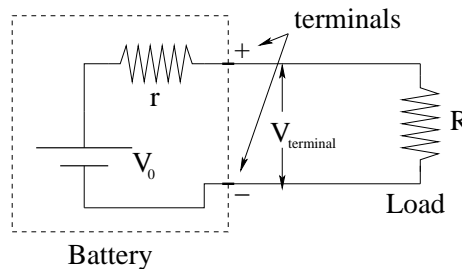


Figure 58: A non-ideal battery in a circuit with a resistive load.

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 delivering 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*.

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 58 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 battery⁵⁶. 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 battery 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:

$$V_0 - Ir - IR = V_0 - I(r + R) = 0 \quad (349)$$

or:

$$I = \frac{V_0}{r + R} \quad (350)$$

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:

$$I_{\max} = \frac{V_0}{r} \quad (351)$$

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 rechargeable 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 :

$$P_r = I^2 r = V_0^2 \frac{r}{(r + R)^2} \quad (352)$$

and

$$P_R = I^2 R = V_0^2 \frac{R}{(r + R)^2} \quad (353)$$

The sum of these 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.

⁵⁶This 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.

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 type, as well as some of the caveats – places where things one might try don't advance you towards the solution.

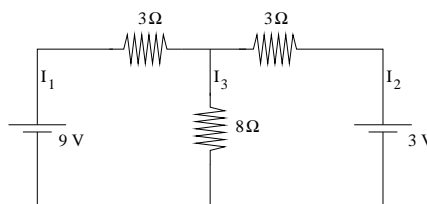


Figure 59: Use Kirchhoff's Rules to find the three unknown currents: I_1, I_2, I_3 .

In figure 59, 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:

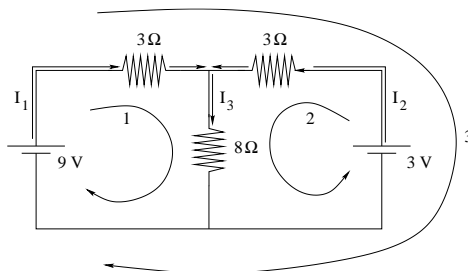


Figure 60: Note the loops and current directions identified on the figure.

Now let us write (and identify) all four equations that we can obtain from Kirchhoff's Rules in this problem:

$$9 - 3I_1 - 8I_3 = 0 \quad (\text{loop 1}) \quad (354)$$

$$3 - 3I_2 - 8I_3 = 0 \quad (\text{loop 2}) \quad (355)$$

$$9 - 3I_1 + 3I_2 - 3 = 0 \quad (\text{loop 3}) \quad (356)$$

$$I_1 + I_2 - I_3 = 0 \quad (\text{junction}) \quad (357)$$

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:

$$3I_1 + 8I_3 = 9 \quad (358)$$

$$3I_2 + 8I_3 = 3 \quad (359)$$

$$I_1 + I_2 = I_3 \quad (360)$$

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:

$$3I_1 + 8I_1 + 8I_2 = 9 \quad (361)$$

$$3I_2 + 8I_1 + 8I_2 = 3 \quad (362)$$

or

$$11I_1 + 8I_2 = 9 \quad (363)$$

$$8I_1 + 11I_2 = 3 \quad (364)$$

If we multiply the top equation by 11 and the bottom equation by 8, we get:

$$121I_1 + 88I_2 = 99 \quad (365)$$

$$64I_1 + 88I_2 = 24 \quad (366)$$

If we subtract the second equation from the first, we get:

$$57I_1 = 75 \quad (367)$$

or

$$I_1 = \frac{75}{57} = 1.316 \quad (368)$$

(in Amperes). We substitute this back into:

$$11\frac{75}{57} + 8I_2 = 9 \quad (369)$$

so

$$I_2 = (9 - 11\frac{75}{57})/8 = -0.785 \quad (370)$$

Finally

$$I_3 = I_1 + I_2 = 1.316 - 0.785 = 0.531 \quad (371)$$

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

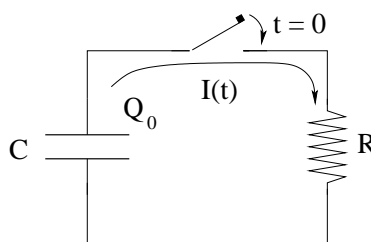


Figure 61: The capacitor C is initially charged to Q_0 . At $t = 0$ the switch is closed and it discharges through the resistor, building up a current $I(t)$.

The capacitor in figure 61 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:

$$\frac{Q}{C} - IR = 0 \quad (372)$$

The current and charge are not independent. The current is, in fact, the rate at which the charge on the capacitor decreases:

$$I = -\frac{dQ}{dt} \quad (373)$$

If we substitute this relation into Kirchhoff's loop rule, divide by R , and rearrange, we get the following equation of motion for Q :

$$\frac{dQ}{dt} + \frac{Q}{RC} = 0 \quad (374)$$

This is a *first order, linear, homogeneous, ordinary differential equation*, in fact the equation for *exponential decay*. It can easily be solved by direct integration. The solution proceeds as follows. Rearrange the equation as follows:

$$\frac{dQ}{dt} = -\frac{Q}{RC} \quad (375)$$

Multiply through by dt , divide through by Q , to get:

$$\frac{dQ}{Q} = -\frac{dt}{RC} \quad (376)$$

Integrate both sides (indefinite integral on the right):

$$\ln(Q) = -\frac{t}{RC} + A \quad (377)$$

(where A is the constant of integration). To get Q , we exponentiate both sides:

$$Q(t) = e^{\ln(Q)} = e^{-\frac{t}{RC} + A} = e^A e^{-t/RC} \quad (378)$$

Finally, we set the constant of integration from the initial conditions, so that $Q(0) = Q_0$:

$$Q(t) = Q_0 e^{-t/RC} \quad (379)$$

From this we can easily find the other quantities mentioned above:

$$I(t) = -\frac{dQ}{dt} = \frac{Q_0}{RC} e^{-t/RC} \quad (380)$$

$$V_C(t) = \frac{Q}{C} = \frac{Q_0}{C} e^{-t/RC} = V_0 e^{-t/RC} \quad (381)$$

$$V_R(t) = -I(t)R = -\frac{Q_0}{C} e^{-t/RC} \quad (382)$$

$$\begin{aligned} P_C(t) &= V_C(t)I(t) = \frac{Q_0}{C} e^{-t/RC} \frac{Q_0}{RC} e^{-t/RC} \\ &= \frac{Q_0^2}{RC^2} e^{-2t/RC} \end{aligned} \quad (383)$$

$$\begin{aligned} P_R(t) &= V_R(t)I(t) = -\frac{Q_0}{C} e^{-t/RC} \frac{Q_0}{RC} e^{-t/RC} \\ &= -\frac{Q_0^2}{RC^2} e^{-2t/RC} \end{aligned} \quad (384)$$

$$(385)$$

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?

$$\begin{aligned} U_R &= \left| \int_0^\infty P_R(t) dt \right| \\ &= \frac{Q_0^2}{RC^2} \int_0^\infty e^{-2t/RC} dt \\ &= -\frac{Q_0^2}{2C} \int_0^\infty e^{-2t/RC} \frac{-2 dt}{RC} \\ &= -\frac{Q_0^2}{2C} e^{-2t/RC} \Big|_0^\infty \\ &= \frac{Q_0^2}{2C} \end{aligned} \quad (386)$$

$$(387)$$

which just *happens* to be the total energy initially on the capacitor:

$$U_C = \frac{1}{2} \frac{Q_0^2}{C} \quad (388)$$

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:

$$\tau = RC \quad (389)$$

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

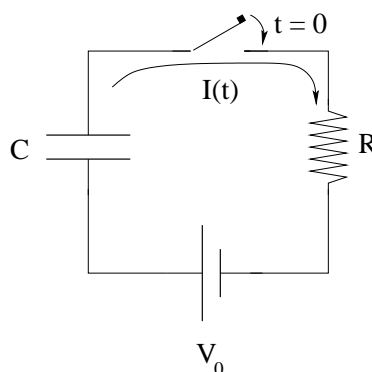


Figure 62: An initially uncharged capacitor being charged through a resistor by a battery with a fixed voltage V_0 .

In figure 62 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:

$$V_0 - \frac{Q}{C} - IR = 0 \quad (390)$$

The current is now the rate at which the charge on the capacitor *increases*:

$$I = +\frac{dQ}{dt} \quad (391)$$

Substituting as before and rearranging, we get:

$$\frac{dQ}{dt} + \frac{Q}{RC} = \frac{V_0}{R} \quad (392)$$

This is a *first order, linear, inhomogeneous, ordinary differential equation*, in fact the equation for *exponential growth*. It, too, can easily be solved by direct integration:

$$\frac{dQ}{dt} = -\frac{Q}{RC} + \frac{V_0}{R} \quad (393)$$

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!*:

$$\frac{dQ}{dt} = -\frac{1}{RC} (Q - CV_0) \quad (394)$$

Now multiply through by dt , divide through by $Q - CV_0$:

$$\frac{dQ}{Q - CV_0} = -\frac{dt}{RC} \quad (395)$$

and integrate both sides (indefinite integral on the right) to get:

$$\ln(Q - CV_0) = -\frac{t}{RC} + A \quad (396)$$

(where A is the constant of integration). As before, to get Q , we exponentiate both sides:

$$Q(t) - CV_0 = e^{\ln(Q - CV_0)} = e^{-\frac{t}{RC} + A} = e^A e^{-t/RC} \quad (397)$$

Finally, we solve for $Q(t)$:

$$Q(t) = CV_0 + e^A e^{-t/RC} CV_0 + B e^{-t/RC} \quad (398)$$

and set the constant of integration $B = e^A$ (the exponential of an unknown constant is still an unknown constant⁵⁷) from the initial conditions, so that $Q(0) = 0$. Our final answer is:

$$Q(t) = CV_0 \left(1 - e^{-t/RC}\right) \quad (399)$$

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 $t \rightarrow \infty$ the total energy provided to the circuit by the voltage equals the total energy stored in the capacitor in the end 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

⁵⁷At 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...

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 teeny 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*⁵⁸ 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.

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 *electrodynamic* model. It is time to move on from the electrostatic field to the next major piece of the electromagnetic puzzle: The magnetic field.

⁵⁸Wikipedia: <http://www.wikipedia.org/wiki/tetrodotoxin>. Found in pufferfish and blue-ringed octupi, for the marine biology crowd.

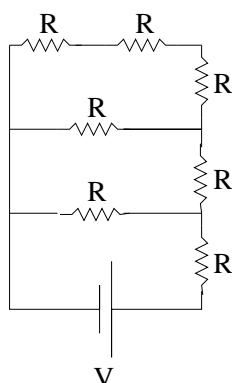
Homework for Week 5

Problem 1.

Physics Concepts

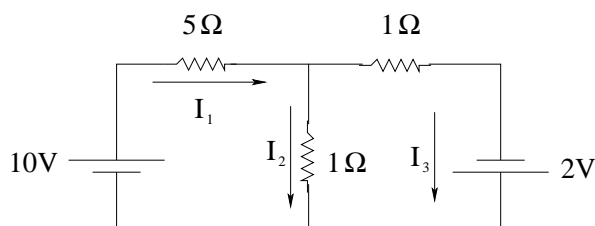
Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



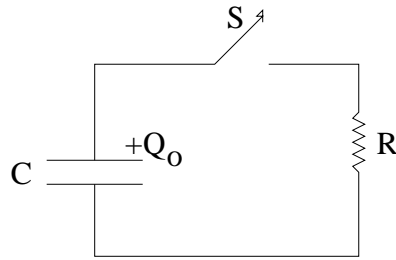
Find the current through each resistor with a voltage V is placed across the resistance network as shown to the left. Note that all of the resistances R are equal. You'll basically need to use the series and parallel rules for adding resistances several times, as well as Ohm's Law and Kirchhoff's junction rule. (Hint: You may find it useful to imagine $V = 18$ volts and $R = 1$ ohm. This makes the numbers easy.)

Problem 3.



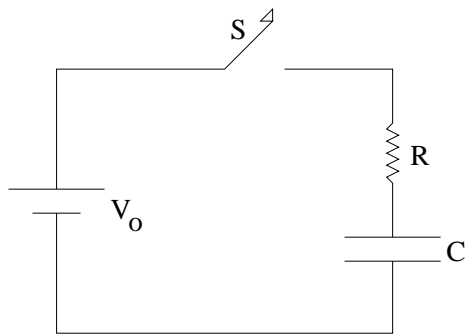
Find the currents I_1 , I_2 , and I_3 in the circuit above.

Problem 4.

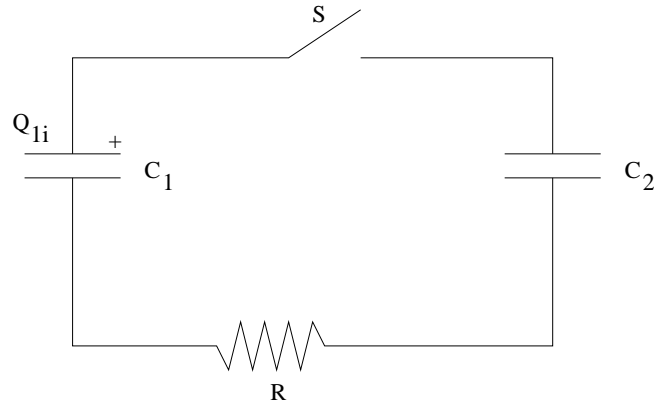


Suppose switch S is closed at time $t = 0$ when the charge on the capacitor is Q_0 . Find $Q(t)$, $I(t)$, $V_C(t)$ and $V_R(t)$ in the circuit above. Find the power delivered to the resistor as a function of time and show that its integral from 0 to ∞ equals the initial energy stored on the capacitor (verifying energy conservation for this circuit).

Problem 5.



Suppose switch S is closed at time $t = 0$ when the charge on the capacitor is $Q_0 = 0$. Find $Q_C(t)$, $I(t)$, $V_C(t)$ and $V_R(t)$ in the circuit above. Find the power delivered to the circuit as a function of time and show that it equals the sum of the power being burned in the resistor plus the power that is charging the capacitor (verifying energy conservation for this circuit).

Problem 6.

A pair of capacitors C_1 and C_2 is connected as shown, with a resistance R in between them. Initially, the first capacitor carries a total charge Q_{1i} and the second one is uncharged, $Q_{2i} = 0$. At $t = 0$ the switch is closed. Find:

- The equilibrium ($t = \infty$) charges on the two capacitors, Q_{1f} and Q_{2f} .
- Using Kirchoff's laws for this arrangement, find the time constant for the equilibration process. Note that you do NOT have to solve the DE, just formulate it with dt and some arrangement of R , C_1 , and C_2 on the other side.
- (Regular) GUESS what the solution to the DE 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.
- (Advanced) Solve the DE (it is integrable, although a bit messy) for $Q_1(t)$ and $Q_2(t)$. It's probably best to solve for just one, and then use conservation of charge to find the other, right?

Advanced Problem 7.

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 – your hint is to think about *current* flowing into and out of this network through probes placed at the junctions and superposition and symmetry. Once you get the square lattice, think about infinite triangular lattices or infinite cubic lattices in 3d.

IV: Magnetostatics

Week 6: Moving Charges and Magnetic Force

(Est 2/13-2/18)

- A charge moving through space is observed to deflect according to the rule:

$$\vec{F} = q(\vec{v} \times \vec{B}) \quad (400)$$

which we use to *define* the magnetic field \vec{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:

$$d\vec{F} = I(d\vec{\ell} \times \vec{B}) \quad (401)$$

for a small (differential) segment of wire carrying a current I in a magnetic field vB . Magnetic fields exert forces on current carrying wires.

- Motion of a point charge in the plane perpendicular to a uniform magnetic field is therefore *circular*:

$$|\vec{F}| = qvB = \frac{mv^2}{r} \quad (402)$$

(Newton's second law plus definition of centripetal acceleration). It has an angular velocity given by:

$$\omega_{\text{cyclotron}} = \frac{qB}{m} \quad (403)$$

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:

$$\vec{m} = NIA\hat{n} \quad (404)$$

where N is the number of turns, I is the current, A is the area, and \hat{n} is the right-handed normal to the plane of the loop.

- The *torque* on a magnetic dipole in a uniform magnetic field is:

$$\vec{\tau} = \vec{m} \times \vec{B} \quad (405)$$

Associated with this are its potential energy:

$$U = -\vec{m} \cdot \vec{B} \quad (406)$$

and its force in a *non*-uniform magnetic field:

$$\vec{F} = -\vec{\nabla}U = \vec{\nabla}(\vec{m} \cdot \vec{B}) \quad (407)$$

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 hesitate*, 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*⁵⁹ 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:

$$\vec{F} = q(\vec{v} \times \vec{B}) \quad (408)$$

Ooo! That pesky *cross product* rears its ugly⁶⁰ 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 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.

Example 6.2.1: A Charged Particle Moving in a Uniform Magnetic Field

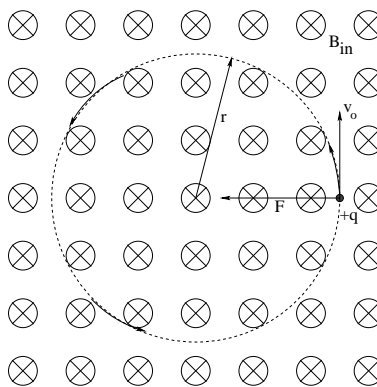


Figure 63: A charge particle with velocity perpendicular to a uniform magnetic field moves in a circle.

In figure(63 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.

⁵⁹Wikipedia: 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!
⁶⁰To 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!

The force \vec{F} acting on this charge is:

$$\vec{F} = q(\vec{v}_0 \times \vec{B}_0) \quad (409)$$

which has magnitude

$$F = qv_0B_0 \quad (410)$$

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:

$$F_r = qv_0B_0 = \frac{mv_0^2}{r} \quad (411)$$

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:

$$\omega_{\text{cyclotron}} = \frac{v_0}{r} = \frac{qB_0}{m} \quad (412)$$

Note well that this 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

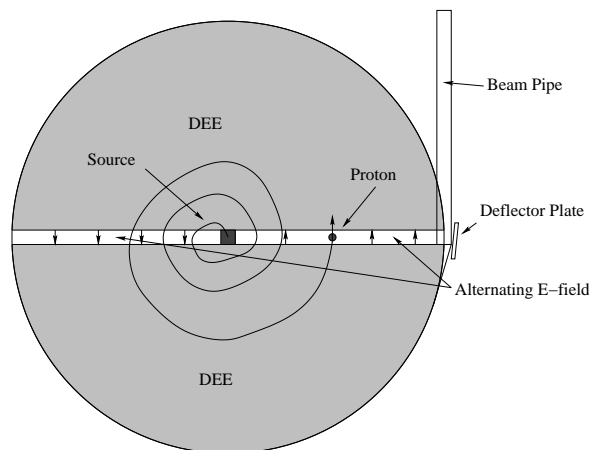


Figure 64: The schematic layout of a cyclotron. The electric field/potential difference between the “Dees” of the cyclotron oscillates with the same period as the period of the cyclotron frequency of the particles moving in the field, so that it always pushes in the direction that speeds it up.

In figure 64 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 accelerating 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 trail, 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

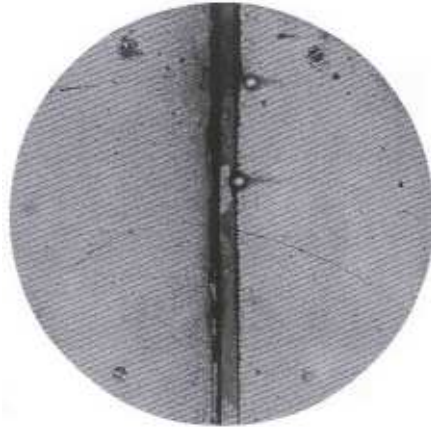


Figure 65: The first photograph of a positron ever taken in a cloud chamber. Note the curvature carefully. Which way is the particle travelling while slowing down? What direction does the magnetic field in the chamber point?

to the charge and mass by:

$$r = \frac{mv_0}{qB_0} \quad (413)$$

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 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*.

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

$$F_B = qvB \quad (414)$$

up. The electric force, however, is

$$F_E = qE \quad (415)$$

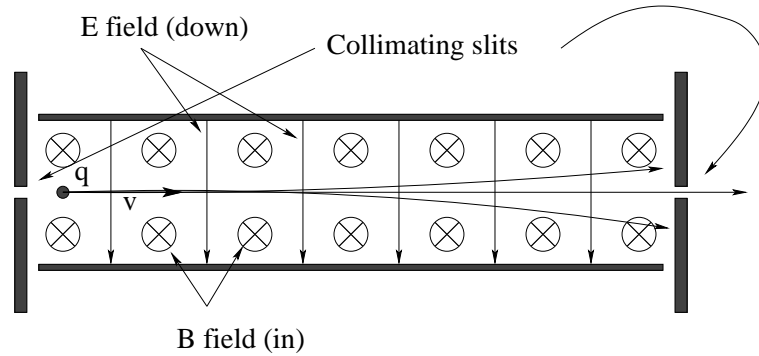


Figure 66: A region of crossed fields functions as a velocity selector; only particles with just the right velocity pass through undeflected.

down. The net force on the particle is *zero* when:

$$F_B = qvB = qE = F_E \quad (416)$$

or when the particle happens to have the velocity

$$v = \frac{E}{B} \quad (417)$$

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

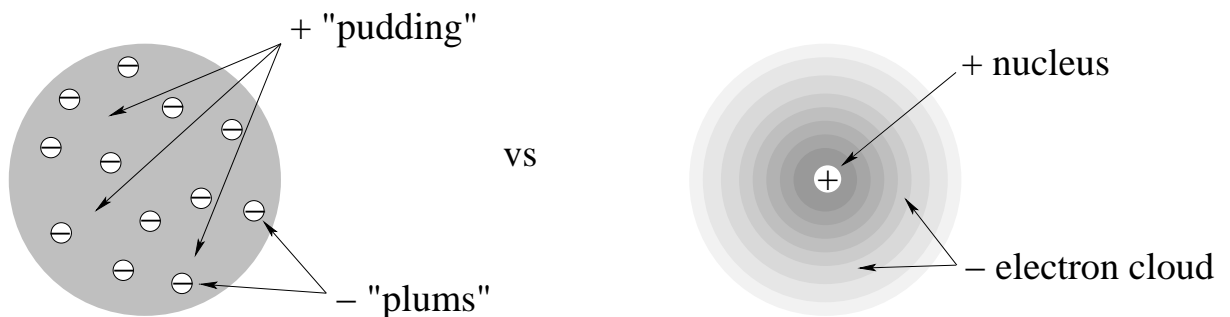


Figure 67: The “plum pudding model” that prevailed in 1897 on left, along with a more accurate representation of the current atomic model – a massive nucleus surrounded by a quantum “pudding” (the electron cloud).

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 67 (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 68 above) for which he received the *first* Nobel Prize in physics in 1901.



Figure 68: 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.

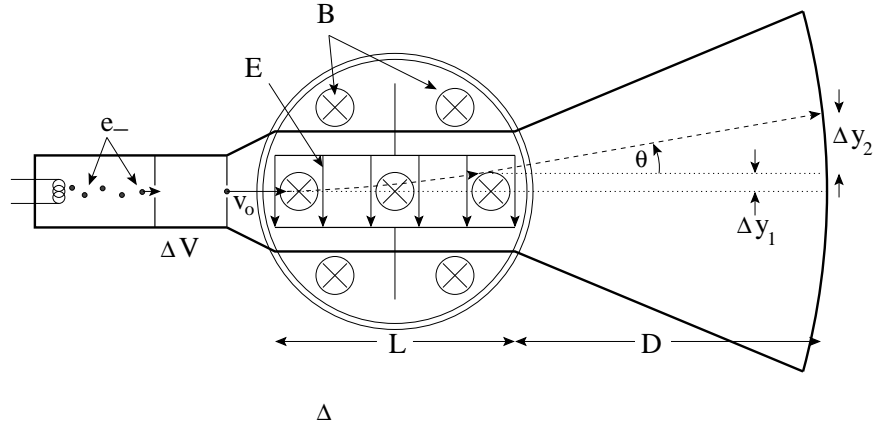


Figure 69: Joseph John Thomson's apparatus for measuring the ratio of the charge on the electron to its mass (improved by the addition of a magnetic velocity selector). This was a critical experiment in determining the structure of the atom, for which Thomson received the Nobel Prize (only the sixth 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 two 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:

$$F_x = 0 \quad (418)$$

$$F_y = eE = ma_y \quad (419)$$

from which we find (using 2D kinematics from the first semester – the problem is *identical* to analyzing trajectories with a constant gravitational acceleration):

$$x(t) = v_0 t \quad (420)$$

$$v_x(t) = v_0 \quad (421)$$

$$y(t) = \frac{1}{2} a_y t^2 = \frac{eE}{2m} t^2 \quad (422)$$

$$v_y(t) = a_y t = \frac{eE}{m} t \quad (423)$$

We can easily find the *time* the electron is between the plates:

$$t_1 = \frac{L}{v_0} \quad (424)$$

from $x(t)$. Substituting this into the last two equations, we find that as it emerges from between the plates:

$$\Delta y_1 = \frac{eE}{2m} t_1^2 = \frac{eEL^2}{2mv_0^2} \quad (425)$$

and

$$v_y = \frac{eE}{m} t_1 = \frac{eEL}{mv_0} \quad (426)$$

From our knowledge of v_x and v_y when the particle emerges, we can find:

$$\tan(\theta) = \frac{v_y}{v_x} = \frac{eEL}{mv_0^2} \quad (427)$$

which lets us easily determine:

$$\Delta y_2 = D \tan(\theta) = \frac{eELD}{mv_0^2}. \quad (428)$$

Now we can relate the *measured* total y deflection to the known values of L , D , E , and our *estimated* v_0 :

$$\begin{aligned} y_{\text{tot}} &= \Delta y_1 + \Delta y_2 \\ &= \frac{eEL^2}{2mv_0^2} + \frac{eELD}{mv_0^2} \\ &= \frac{eEL}{mv_0^2} \left(\frac{L}{2} + D \right) \\ &= \frac{e}{m} \frac{EL}{v_0^2} \left(\frac{L}{2} + D \right) \end{aligned} \quad (429)$$

Inverting this last relation we find:

$$\frac{e}{m} = \frac{y_{\text{tot}} v_0^2}{EL \left(\frac{L}{2} + D \right)} \quad (430)$$

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:

$$v_0 = \frac{E}{B} \quad (431)$$

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:

$$\frac{e}{m} = \frac{y_{\text{tot}} \left(\frac{E}{B} \right)^2}{EL \left(\frac{L}{2} + D \right)} = \frac{y_{\text{tot}} E}{B^2 L \left(\frac{L}{2} + D \right)} \quad (432)$$

This permits a measurement that is as accurate as one's knowledge of y_{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

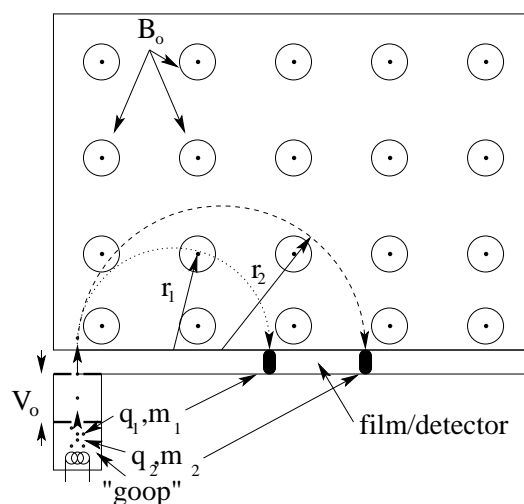


Figure 70: The Mass Spectrometer uses a region with a uniform magnetic field to create a *spectrum* of particles that collide with a film or other detector matrix in places that indicate the radius of the circle they are bent in by the field. This, in turn, is related to the ratio of q/m for the particle, and by assuming a charge that is a low integral multiple of e one can determine the mass.

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, \dots and (associated) charge q_1, q_2, q_3, \dots are then accelerated by a potential difference to an energy (respectively) of $q_1 V_0, q_2 V_0, q_3 V_0, \dots$, passed through a pair of collimating slits as 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:

$$\frac{1}{2}mv^2 = qV_0 \quad (433)$$

in the beam entering the magnetic field. It therefore has a velocity⁶¹:

$$v = \sqrt{\frac{2qV_0}{m}} \quad (434)$$

and experiences a centripetal magnetic force (that causes it to move in a circle of radius r) of:

$$F_r = qvB_0 = m\frac{v^2}{r} \quad (435)$$

so as usual:

$$\frac{v}{r} = \frac{q}{m}B_0 \quad (436)$$

If we solve for the radius r of its half-orbit to the film/detector, we get:

$$r = \frac{v}{B_0} \frac{m}{q} \quad (437)$$

Substituting for v :

$$r = \frac{\sqrt{\frac{2qV_0}{m}}}{B_0} \frac{m}{q} = \sqrt{\frac{2mV_0}{qB_0^2}} = \sqrt{\frac{m}{q}} \frac{\sqrt{2V_0}}{B_0} \quad (438)$$

Alternatively, since one measures r and wishes to find m (given a good guess for q):

$$m = \frac{r^2 B_0^2}{2V_0} q \quad (439)$$

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.

⁶¹As 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 .

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).

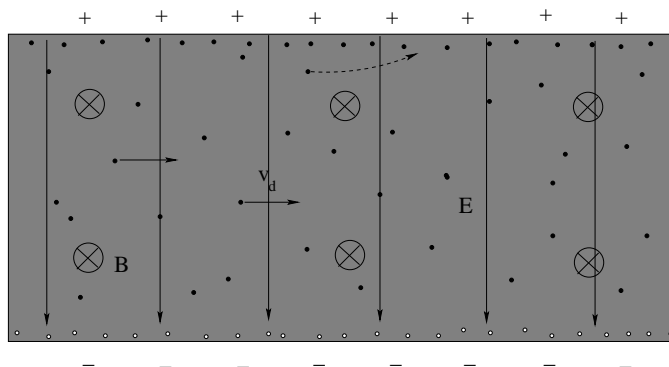


Figure 71: In the Hall Effect, a magnetic field causes the *mobile* charge to accumulate on the upper or lower edge of a conducting, current-carrying strip in a magnetic field. This in turn creates a potential difference across the strip that can easily be measured.

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 71 above, we assume *positive* charge 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:

$$qv_d B = qE \quad (440)$$

or

$$v_d = \frac{E}{B} \quad (441)$$

We also know that:

$$I = nqv_d A = nqA \frac{E}{B} \quad (442)$$

Finally, we know that

$$V = Ew \quad (443)$$

or

$$E = \frac{V}{w} \quad (444)$$

so that

$$I = nqA \frac{V}{Bw} \quad (445)$$

We can then solve for n , the desired density of charge carriers:

$$n = \frac{IBw}{qAV} \quad (446)$$

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:

$$I = nqv_d A = \int \vec{J} \cdot \hat{n} dA \quad (447)$$

(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 $d\ell$:

$$d\vec{F} = nq(Ad\ell)\vec{v}_d \times \vec{B} \quad (448)$$

We now do a clever thing. We’ll collect the nqv_dA magnitudes together and make I , and take the *direction* of \vec{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:

$$d\vec{F} = I(d\vec{\ell} \times \vec{B}) \quad (449)$$

for a small (differential) segment of wire carrying a current I in a magnetic field \vec{B} . 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)

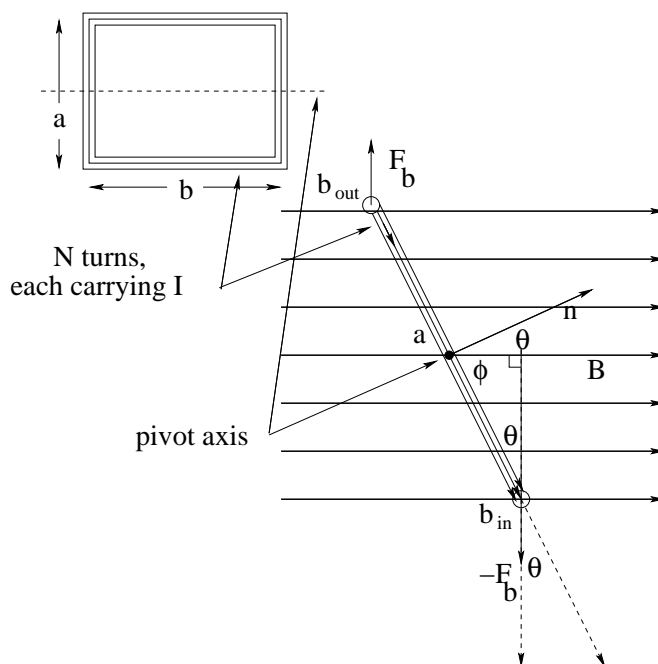


Figure 72: The force and torque on an $a \times b$ rectangular loop of N turns, each carrying current I , in a uniform magnetic field \vec{B} are $\vec{F} = 0$ and $\vec{\tau} = \vec{m} \times \vec{B}$ respectively.

In figure 72 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 pivots 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{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{B} field, have length b , have N turns each carrying I , and hence the magnitude of the force is:

$$F_b = N I b B \quad (450)$$

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:

$$F_a = |N I \vec{a} \times \vec{B}| = N I a B \sin(\phi) \quad (451)$$

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

$$\vec{F}_{\text{tot}} = (F_b - F_b)\hat{y} + (F_a - F_a)\hat{z} = 0 \quad (452)$$

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 ($\vec{F} = 0$) 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 torque 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 $\vec{\tau} = \vec{r} \times \vec{F}$ where \vec{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:

$$\tau_b = \frac{a}{2} F_b \sin(\theta) \quad (453)$$

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:

$$\tau = a F_b \sin(\theta) = N I (a b) B \sin(\theta) \quad (454)$$

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{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:

$$\vec{m} = NI(ab)\hat{n} \quad (455)$$

in which case the torque takes the familiar form:

$$\vec{\tau} = \vec{m} \times \vec{B} \quad (456)$$

which looks *just like* the torque on an *electric* dipole, $\vec{\tau} = \vec{p} \times \vec{E}$! 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

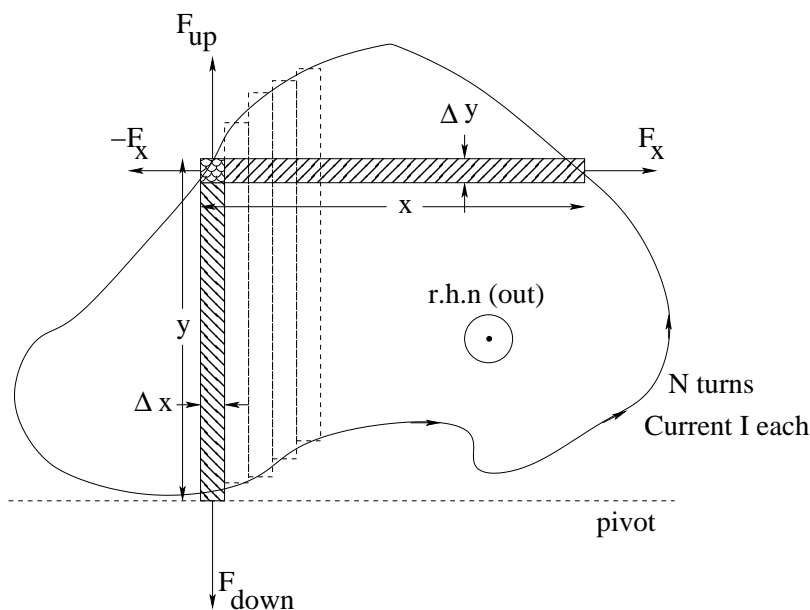


Figure 73: Arbitrary plane loop of current can be broken into small pieces that are aligned with or perpendicular to torque axis.

In figure 73 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{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{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:

$$\Delta\tau = y\Delta x NIB \sin(\theta) \quad (457)$$

Summing over all of the strips of width Δx , the total torque on this plane loop is thus:

$$\tau = NI \lim_{\Delta x \rightarrow 0} \left(\sum y(x) \Delta x \right) B \sin(\theta) = NI \left(\int y(x) dx \right) B \sin(\theta) = NIAB \sin(\theta) \quad (458)$$

or (including the vector direction from the right-hand-rule applied both to the torque and the right handed normal to the loop):

$$\vec{\tau} = \vec{m} \times \vec{B} \quad (459)$$

with

$$\vec{m} = NIA\hat{n} \quad (460)$$

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.

0.1 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

$$U = -\vec{m} \cdot \vec{B} \quad (461)$$

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:

$$F_x = - \frac{dU}{dx} \quad (462)$$

(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 torques. It is quite strong:

Quantity	Electric Dipole	Magnetic Dipole
Dipole Moment	$\vec{p} = q\vec{\ell}$	$\vec{m} = NIA\hat{n}$
Force in Uniform Field	$\vec{F} = 0$	$\vec{F} = 0$
Torque in Uniform Field	$\vec{\tau} = \vec{p} \times \vec{E}$	$\vec{\tau} = \vec{m} \times \vec{B}$
Potential Energy	$U = -\vec{p} \cdot \vec{E}$	$U = -\vec{m} \cdot \vec{B}$
Force in Non-Uniform Field	$\vec{F} = -\vec{\nabla}U$	$\vec{F} = -\vec{\nabla}U$

Table 3: Similarity of results for the electric and magnetic dipoles in (or later, as the source of) their respective fields.

Example 6.3.3: 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.

There is one more generic distribution of moving charge 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).

If we imagine the proton to be a uniform ball of charge with radius R , total (uniformly distributed) charge $Q = e$, (uniformly distributed) mass M , spinning about some axis through its center at an angular velocity $\vec{\omega}$ so that:

$$\vec{L} = I\vec{\omega} = \left(\frac{2}{5}MR^2\right)\vec{\omega} \quad (463)$$

then it is clear that the proton will *also* have a *magnetic moment* parallel to \vec{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

$$\vec{m} = \frac{Q}{2M}\vec{L} = \mu\vec{L} \quad (464)$$

for *any* symmetric spinning particle with identically distributed charge and mass, where I have defined the ratio:

$$\mu = \frac{Q}{2M} \quad (465)$$

as the classical equivalent of the “Bohr magneton” in the quantum physics of the electron⁶². Let us understand this.

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 74

⁶²The 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.

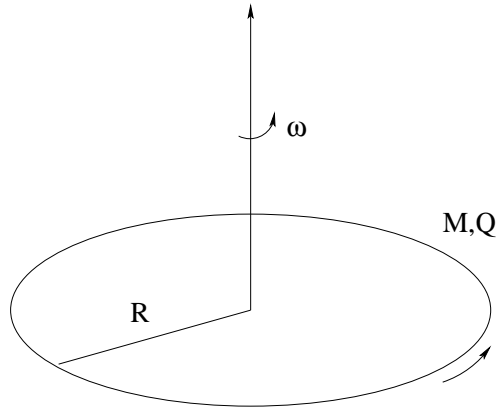


Figure 74: A rotating ring of charge with mass M , radius R , and charge Q has a magnetic moment of $\vec{m} = Q/2M(MR^2\vec{\omega}) = \mu\vec{L}$.

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:

$$I = \frac{Q}{T} = \frac{Q\omega}{2\pi} \quad (466)$$

The magnetic moment of the ring in the (right handed) z -direction is thus just:

$$m_z = IA = \frac{Q\omega}{2\pi}\pi R^2 = \frac{Q\omega}{2}R^2 \quad (467)$$

If we multiple the expression on the right by $\frac{M}{M}$ (one!) and rearrange the terms, we get:

$$m_z = \frac{Q}{2M}(MR^2\omega) = \mu L_z \quad (468)$$

using $L_z = MR^2\omega$ for a ring of mass M rotating symmetrically about the z -axis.

That was almost too easy! 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 $dq = (2\pi r dr)\sigma$ where $\sigma = Q/\pi R^2$ is the surface charge density of the uniformly distributed charge. $dI = dq\frac{\omega}{2\pi}$ as before. The area inside the ring is $A = \pi r^2$. Thus:

$$dm_z = dIA = 2\pi r dr \frac{Q\pi r^2}{\pi R^2} \frac{\omega}{2\pi} = \frac{Q}{R^2}\omega r^3 dr \quad (469)$$

and integrating from 0 to R we find:

$$m_z = \int_0^R \frac{Q}{R^2}\omega r^3 dr = \frac{Q}{4R^2}R^4\omega = \frac{Q}{4}R^2\omega \quad (470)$$

One again we multiply by $\frac{M}{M}$, do some rearrangement, and:

$$m_z = \frac{Q}{4M}MR^2\omega = \frac{Q}{2M}\left(\frac{1}{2}MR^2\omega\right) = \mu L_z \quad (471)$$

Part of your homework for this week will be to re-prove these two cases and several others to show that this result is quite general. If you are an advanced student (e.g. a physics major) you may be asked to prove that this is a *general* result for any figure with enough symmetry about the axis of rotation so that $\vec{\omega} \parallel \vec{L}$. 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. 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:

$$\vec{m}_p = \mu_p \vec{L}_p \quad (472)$$

where $m\mu_p = \frac{e}{2m_p}$, the magnetic moment of a classical electron should be the same with $\mu_e = \frac{e}{2m_e}$ and so on. This result works adequately in the quantum case 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.

Example 6.3.4: The Precession of Magnetic Moments: Magnetic Resonance

In figure 75 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* vL 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 its magnetic moment can generally enough be written:

$$\vec{m} = \frac{e}{2m_p} \vec{L} = \mu_p \vec{L} \quad (473)$$

where $e = 1.6 \times 10^{-19}$ Coulombs is its charge, $m_p = 1.67 \times 10^{-27}$ kilograms is its mass (in SI units).

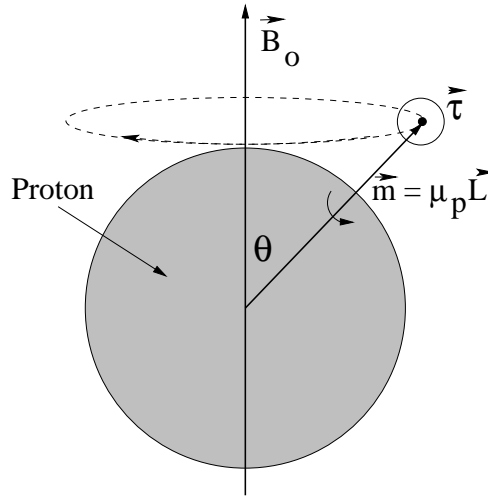


Figure 75: A rotating proton with a magnetic moment $\vec{m} = \mu_p \vec{L}$ aligned with its rotation axis *precesses* around an applied magnetic field \vec{B}_0 with precession frequency $\omega_p = \mu B_0$ independent of the particular angle θ between \vec{m} and \vec{B}_0 . Note that $\mu_p = \frac{e}{2m_p}$ from the previous section.

The magnetic field exerts a *torque* $\vec{\tau}$ on the magnetic dipole of the proton, out of the page at the particular instant drawn, that is given by:

$$\vec{\tau} = \vec{m} \times \vec{B}_0 \quad (474)$$

or (using the fundamental definition of the torque as the time rate of change of the angular momentum):

$$\frac{d\vec{L}}{dt} = \mu_p (\vec{L} \times \vec{B}_0) \quad (475)$$

This is a very important first order, linear, homogeneous, ordinary differential equation. It describes *precession*, a common phenomenon in physics, especially when considering motion in rotating frames. It *does* have exponential forms for its solution, but we will at first skip over a detailed treatment of this approach and “solve” it more intuitively, graphically and algebraically.

The magnitude of the torque is given by:

$$\tau = \mu_p L B_0 \sin(\theta) \approx \frac{\Delta L}{\Delta t} \quad (476)$$

The direction is out of the page. If we break \vec{L} up into two components, one parallel and one perpendicular to \vec{B}_0 as shown in the simplified figure ??, we can easily see that the torque only comes from, and only changes, the L_\perp component.

The torque is always *perpendicular* to \vec{L}_\perp , and hence 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 76 at an instant when \vec{L}_\perp has both x and y components.

In a short time Δt , the angular momentum changes a small amount ΔL_\perp that is the length of the arc on the L_\perp circle subtended by the angle through which it turns in that time, $\Delta\phi$. That is:

$$\Delta L = \Delta L_\perp = L_\perp \Delta\phi \quad (477)$$

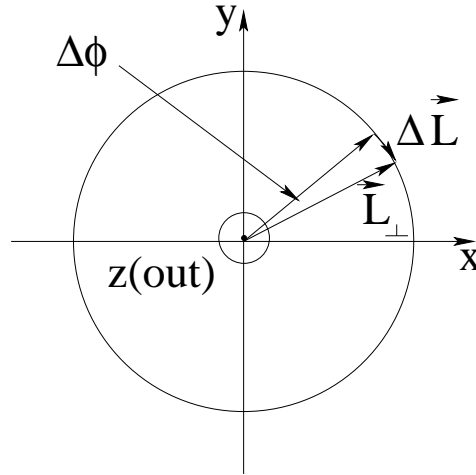


Figure 76: The torque causes L_{\perp} to precess around the z -axis (out of the page as drawn). L_{\perp} moves in a circle, and in a short time Δt it moves through an angle $\Delta\phi$ and hence changes the (vector) angular momentum by $\Delta\vec{L}$ as shown.

or (dividing both sides by Δt):

$$\frac{\Delta L}{\Delta t} = L_{\perp} \frac{\Delta\phi}{\Delta t} \quad (478)$$

This must *also* equal the torque in terms of the field, in the limit that we let $\Delta t \rightarrow dt$:

$$\tau = \frac{dL}{dt} = \mu_p L_{\perp} B_0 = L_{\perp} \frac{d\phi}{dt} = L_{\perp} \omega_p \quad (479)$$

where we used $L_{\perp} = L \sin(\theta)$ and $\omega_p = \frac{d\phi}{dt}$ (the angular precession frequency). Solving for the latter, we find:

$$\omega_p = \mu B_0 \quad (480)$$

independent of the angle θ between \vec{m} and \vec{B}_0 !

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 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 optional exercise for the interested (non-major) student or a required problem for physics/engineering/math majors in the homework. 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!

6.4: 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.

MRI is primarily used medically to map the density of *hydrogen nuclei* – protons – in the human body. 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.

From our discussion above, we *classically* expect the magnetic dipole moment of an isolated proton to be given by an expression such as:

$$\vec{m}_p = \frac{e}{2m_p} \vec{S} \quad (481)$$

There are several problems with this, however. The proton is not, in fact, a homogenous ball of spinning mass and charge. It is a composite particle made up of three quarks, and *they* have spin angular momentum \vec{S} as shown (a purely quantum mechanical kind of angular momentum), not the classical “orbital” angular momentum described by \vec{L} . It is therefore simpler to describe the actual ratio of the magnetic dipole moment of a proton to its actual spin angular momentum as a semi-empirical parameter and define:

$$\vec{m}_p = \gamma \vec{S} \quad (482)$$

where γ is the so-called *gyromagnetic ratio* for the proton and where we will continue to use \vec{L} as the symbol for its nuclear spin angular momentum. A concise discussion of this and accepted values for γ are available on Wikipedia: http://www.wikipedia.org/wiki/Proton_magnetic_moment, although 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 $\sqrt{s(s+1)}\hbar = \sqrt{3}/2\hbar$ for a spin- $\frac{1}{2}$ particle.

This is where the subtleties of quantum mechanics start to kick in. In the absence of any strong magnetic field, the total spin angular momentum of any collection of protons will be approximately zero. The spins will *not* necessarily be aligned with any given z -axis direction (which is, after all, arbitrary). We can pick a z -axis and try to describe the spin state of any given proton in terms of eigenstates of S_z relative to that direction, but if we do all we will get is that spins are in *completely*

random superpositions of those states, ones that are as likely to be eigenstates of S_x or S_y (or the spin resolved relative to some random z -axis direction $S_{z'}$) as in pure S_z eigenstates.

We are thus forced into using a *statistical* description of the state of the protons. Basically, their spin angular momentum (and magnetic moments) are equally likely to point in all possible directions. This is not saying that each proton doesn't *have* a particular spin direction at any given time, only that we don't know what it is.

If we put this collection of more or less isolated spins into a strong external magnetic field, however, the situation changes. Now the field itself *defines* a suitable z -direction. Furthermore, we know that the *energy* of the protons changes. The potential energy of each proton in a strong external magnetic field $\vec{B} = B_0\hat{z}$ is given by:

$$U = -\vec{m}_p \cdot \vec{B} = -m_z B_0 \quad (483)$$

This energy is minimized when the spin is *aligned* with the external field, and is maximized when it is *antialigned* with the external field. At any finite temperature, the lower energy of spins when they are aligned makes this state *more probable* than spins aligned the other way. In equilibrium one then expects to have *more* spins aligned with the field than anti-aligned and the system develops a finite total magnetic moment aligned with the field (where in the absence of the field, its total magnetic moment is within random fluctuations of zero).

This state of affairs is not realized instantly, however. If one starts with no field and then “suddenly” turns on a strong field, the system will relax towards the equilibrium state (approximately) *exponentially*, with a characteristic relaxation time that we will call T_1 . During this relaxation, the system *loses electromagnetic energy* via radiation and spin-lattice relaxation processes but these processes are largely incoherent and do not produce a detectable signal.

Homework for Week 6

Problem 1.

Physics Concepts

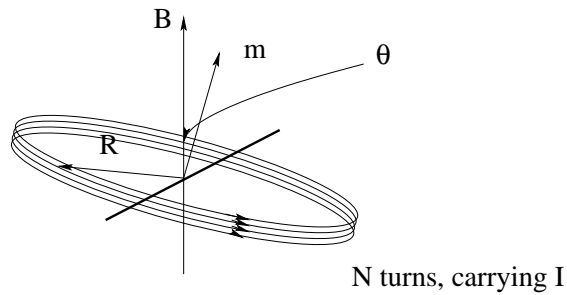
Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.

A particle with mass m and charge q 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 one 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 .

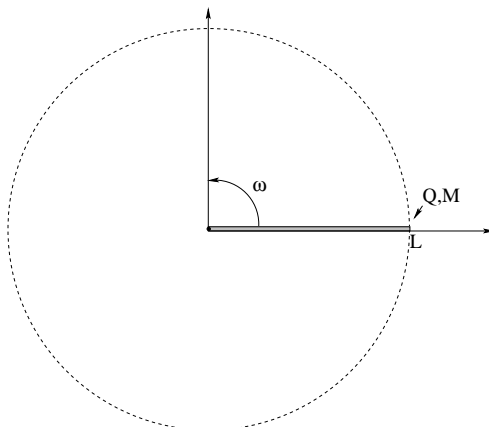
Problem 4.

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 $\vec{B} = B_0 \hat{z}$ so that its magnetic moment makes an initial angle of $\theta \ll \pi$ 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 and will not precess.

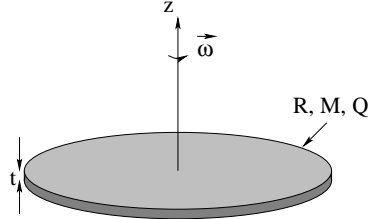
Problem 5.

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 $\vec{B} = B_0 \hat{z}$. Find the frequency ω_p with which the magnetic moment *precesses* around the magnetic field.

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 .

- 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 .
- Integrate this result and find the total magnetic (dipole) moment of the rotating rod m_z
- 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.

- Find the magnetic moment dm_z of this chunk in terms of ρ_q , ω , dV and its coordinates.
- Find the angular momentum dL_z of this chunk in terms of ρ_m , ω , dV and its coordinates.
- 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 $\mu_B = \frac{Q}{2M}$.
- 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.

Using the insight gained from the previous two problems, consider any of the symmetric distributions of charge and mass, where the mass distribution is the same as the charge distribution and where both are “balanced” rotationally. Find a relationship between dI (the moment of inertia of a small chunk of mass dm at a radius r) and dm_z (the magnetic moment of the *same* small chunk of charge dq at the radius r) to show that for all distributions with sufficient (balanced) symmetry that $L_z = I\omega$, $m_z = \frac{Q}{2M} L_z$. This result therefore holds for spheres, cylinders, disks, rods (in a plane), spherical or cylindrical shells, etc.

Advanced Problem 9.

A semi-infinite thin solenoid aligned with (say) the negative z -axis so that the “+” end is at the origin creates a magnetic field that *looks* like that of a point magnetic charge q_m at the origin:

$$\vec{B} = \frac{k_m q_m \hat{r}}{r^2}$$

at points “near” the end and outside of the solenoid itself. Note that $k_m = \mu_0/4\pi = 10^{-7}$ N-m/A² is the magnetic field constant, analogous to k_e for the electric field, and that μ_0 is called the *magnetic permeability*, none of which matters more than algebraically for this problem but which is important next week!

Suppose you take a small bar magnet and place it at $\vec{r} = r\hat{r}$ so its magnetic moment \vec{m} is aligned with \hat{r} . Find the force acting on it (if any).

What would you expect its motion to be if you placed it at the same point so that its moment was *not* initially aligned with the magnetic field?

Advanced Problem 10.

For a presumed e.g. proton in a magnetic field, evaluate:

$$\frac{d\vec{L}}{dt} = \mu_p \vec{L} \times \vec{B} \quad (484)$$

in Cartesian components, assuming $vB = B_0\hat{z}$ and arbitrary $\vec{L} = L_x\hat{x} + L_y\hat{y} + L_z\hat{z}$. Identify and solve the resulting system of equations to 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{L} 's magnitude or direction.

Week 7: Sources of the Magnetic Field

(Est 2/18-2/25)

- No *isolated magnetic monopoles* 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, in spite of an intense experimental search for them. It is probably safe to say that magnetic monopoles are at the very least *rare*.
- We express this (lack of monopoles) by means of *Gauss’s Law for Magnetism*:

$$\oint_S \vec{B} \cdot \hat{n} dA = 4\pi k_m Q_{m,\text{in } S} = \mu_0 \int_{V/S} \rho_m dV = 0 \quad (485)$$

where the magnetic field constant $k_m = 10^{-7}$ tesla-meter/ampere *exactly* (exactly because it defines the coulomb, not the other way around).

- The actual source for magnetic fields (in the absence of monopoles) is *moving charge*. The field produced by a point charge is given by:

$$\vec{B} = k_m \frac{q\vec{v} \times \hat{r}}{r^2} = \frac{\mu_0}{4\pi} \frac{q\vec{v} \times \hat{r}}{r^2} \quad (486)$$

where $\mu_0 = 4\pi \times 10^{-7}$ tesla-meter/ampere is called the *magnetic permeability of free space* and is the magnetic constant analogous to ϵ_0 , the dielectric permittivity of free space.

- If we consider a wire carrying a current $I = nqv_d A$ (where recall v_d is the average drift speed of the charge carriers q), the amount of charge in a small length of wire $d\ell$ is $dq = nqAd\ell$. The field it produces is therefore:

$$\begin{aligned} d\vec{B} &= k_m \frac{dq\vec{v}_d \times \hat{r}}{r^2} \\ d\vec{B} &= k_m \frac{nqAd\ell\vec{v}_d \times \hat{r}}{r^2} \\ d\vec{B} &= k_m \frac{nqv_d A d\vec{\ell} \times \hat{r}}{r^2} \\ d\vec{B} &= k_m \frac{Id\vec{\ell} \times \hat{r}}{r^2} \end{aligned}$$

where $d\vec{\ell}$ is a differential length of the wire with a direction pointing *in the direction of the current*. This:

$$d\vec{B} = k_m \frac{Id\vec{\ell} \times \hat{r}}{r^2} \quad (487)$$

is known as the *Biot-Savart Law* for the magnetic field, and (one way or another) is the way most of the magnetostatic fields we observe in nature come into being.

- The field of a long straight wire carrying a current I is:

$$\vec{B} = \frac{2k_m I}{r} \hat{\phi} \quad (488)$$

where $\hat{\phi}$ curls around the wire in the direction given by the *right hand rule*.

- Learn to use the Biot-Savart law to find the field of a long straight wire, a current carrying loop, and a rotating 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. (See homework)
- With more work than we can do in this course the Biot-Savart Law can be used to prove *Ampere's Law*:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 I_{\text{thru } C} = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \quad (489)$$

This is our *third* Maxwell equation.

- There is a *conceptual error* in Ampere's Law. The current I through an open surface S bounded by a closed curve C is *not invariant* as we vary all possible such surfaces! 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:

$$\frac{dQ}{dt} = - \oint_S \vec{J} \cdot \hat{n} dA \quad (490)$$

you should be able to *deduce* the necessity for *Maxwell's Displacement Current* (which makes the total current invariant). If you can do this on your own without looking and show me the algebra, you get a piece of candy! Sorry, 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 and solve it. Unify a field. You too can have your name on something!

- Learn to use Ampere's Law to find the magnetic field of any cylindrically symmetric current distribution, a (long) solenoid, and a toroidal solenoid. (See homework)
- Useful true fact: We do not usually deduce a *scalar* magnetic potential analogous to the electric potential. Instead you will eventually learn about a *vector* potential that leads to the magnetic field by virtue of differentiation (the curl). Because it is a vector, it is not much easier to evaluate directly than the Biot-Savart law above (it involves doing a very similar but slightly simpler integral). We will therefore skip it altogether in this course.

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. It is time to search for the *sources* of this field.

It is perfectly reasonable to begin our search by saying to ourselves: "Gee, I 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?"

If there is, then we would expect the field of a collection of “magnetic monopolar” charges q_m ⁶³ to be given by:

$$\vec{B}(\vec{r}) = \sum_i \frac{k_m q_{mi}(\vec{r} - \vec{r}_i)}{|\vec{r} - \vec{r}_i|^3} \quad (491)$$

and the magnetic force between a pair of monopoles to be given by:

$$\vec{F}_{12} = \frac{k_m q_{m1} q_{m2}(\vec{r}_1 - \vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|^3} \quad (492)$$

where I’ve introduced a magnetic force constant equivalent to k_e to set the 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*:

$$\oint_S \vec{B} \cdot \hat{n} dA = 4\pi k_m Q_{m,\text{in } S} = \mu_0 \int_{V/S} \rho_m dV \quad (493)$$

proceeding *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. In fact:

$$k_m = 10^{-7} \text{tesla} - \text{meter/ampere} \quad (494)$$

exactly (exactly because it defines the coulomb, not the other way around) as was determined and defined by Ampere in his experiments on magnetism!

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 60 years at this point intense effort has been expended searching experimentally for magnetic monopoles using a variety of ingenious methods.

Alas, no *isolated magnetic monopoles* have been experimentally observed, in spite of an electromagnetic theory that “begs” 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! 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 monopoles by modifying Gauss’s Law of Magnetism to be:

$$\oint_S \vec{B} \cdot \hat{n} dA = 4\pi k_m Q_{m,\text{in } S} = \mu_0 \int_{V/S} \rho_m dV = 0 \quad (495)$$

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. 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 end at a point for electric field lines if the point contained an *electric charge*. Nor can they cross. This leaves only one alternative:

⁶³I 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.

- 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 at first.

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 could *change* Gauss's Law for magnetism (as well as other Maxwell equations) and we'd all have to work a bit harder to learn electrodynamics. But it would very definitely be worth it to be able to understand why charge is quantized.

Magnetic Flux

Although (to be honest) there isn't a lot of point to it, this work would not be complete without a mention of the definition of magnetic flux through a surface S and the SI units associated with magnetic flux. Magnetic flux, it seems, *gets* its own units where electric flux does not, although we haven't missed the lack of units for electric flux and wouldn't miss not knowing the units of magnetic flux *unless* you (sigh) read papers or other textbooks and they refer to them. Since some of you might one day do just that, I suppose we should define them.

The definition of magnetic flux through a surface S is (as should already be clear):

$$\phi_m = \int_S \vec{B} \cdot \hat{n} dA \quad (496)$$

(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, as you prefer.

There. That's done. You may now forget Webers, without *really* forgetting Webers, if you know what I mean. We won't use them, I'm certainly not going to ask you what they are on a test or quiz, but you should probably know the True Fact that they are the SI units of magnetic flux so you can understand what people are talking about who use them in a sentence like "Gosh, my magnetic flux feels like it is up to 2.17×10^{-7} Webers today, no wonder I feel terrible."

Maybe I *should* give you an assignment on this: Go up to a non-physics-educated friend who is very pretentious and puts on intellectual airs and ask him or her if he/she was worried about the rapidly varying webers through the local power grid in the latest solar storm. Enjoy his or her profoundly pained or puzzled look for the precious moment that it lasts... oh yeah, did I do that? *Schooled!*

Note well that while we won't use the *units* by *name*, it will turn out that the *definition* of magnetic flux will be very important to us, both in Gauss's Law for Magnetism (where zero or not, it is important in what it tells us) and in the yet-to-be-learned *Faraday's Law*, our final Maxwell equation. Next week.

7.2: The Magnetic Field of a Point Charge

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. They are for the most part produced by *moving electrical charges*⁶⁴.

Time to go into a laboratory and perform experiments. These experiments are actually rather difficult to do for single charges, so (as we'll see in a moment) the original experiments more commonly involved electrical *currents* made up of many moving charges, but we'll find it a bit more useful to start at the microscopic end and then average to find the macroscopic rule. From them we learn that:

- The magnetic field of a moving (electrical) charge q is proportional to the charge.
- The magnetic field of a moving charge is inversely proportional to the distance from the charge squared, just as was the case of the electrostatic field.
- The *direction* of the magnetic field is that it forms *loops* around the direction of motion of the charge, in the direction the fingers of your right hand curl around your thumb if you line your thumb up along with the direction of motion of the positive charge.
- For any given value of q , v and r , the field strength is proportional to $\sin(\theta)$ where θ is the angle between the direction of \vec{v} and the direction of \vec{r} .

These empirical observations can be summarized in the following formula⁶⁵:

$$\vec{B} = k_m \frac{q (\vec{v} \times \hat{r})}{r^2} = \frac{\mu_0}{4\pi} \frac{q (\vec{v} \times \hat{r})}{r^2} \quad (497)$$

so that the magnetic field has the *magnitude*:

$$B = k_m q v \sin(\theta) \frac{1}{r^2} \quad (498)$$

The geometry of the field lines is shown in figure 77. Note that if you let the thumb of your right hand line up with 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.

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 its value at a point at a lagged time *after* the charge or current appears at any given point.

⁶⁴Reality is slightly more complicated than that, however. For one thing, point-like electrons have a magnetic dipole moment due to their spin even though it is difficult to describe a point-like object as “moving”. For another, as we shall see, 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 moving electrical charge at least at first, and later can visualize and bootstrap an understanding of point-like magnetic dipoles as a kind of limit of rotating macroscopic charge.

⁶⁵This equation was originally obtained by Oliver Heaviside in 1888, long after the Biot-Savart Law for *currents* (next) was discovered, but it is in some sense more fundamental. However, it is also *flawed*. For one thing, it lacks *retardation* – the field emitted from the moving charge propagates at the speed of light and is not instantaneous. This is not a problem with the Biot-Savart Law because it is a *magnetostatic* law valid for more or less continuous and steady currents. It is worth keeping in mind while reading this chapter that with the exception of Gauss's Law for Magnetism, every fundamental equation taught herein is *not quite true* and is presented as they are in elementary courses to help bootstrap Maxwell's Equations as they are eventually (correctly) written.

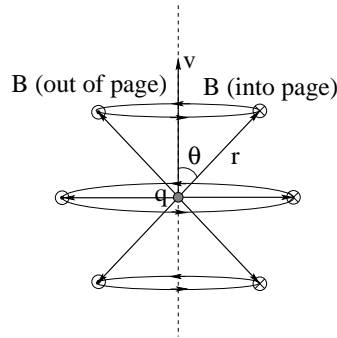


Figure 77: The geometry of the magnetic field lines going in circles *around* the (dotted) line of motion of the charge in the right-handed sense. Note that the direction of $\vec{v} \times \vec{r}$ is into the paper on the right, out of the paper on the left, for \vec{v} and any of the \vec{r} vectors shown.

For the field “near” a more or less stationary (slowly moving) electric charge this didn’t matter much, but for a moving charge (where there is no magnetic field at all without the motion) it *starts* to matter, and indeed advanced students may wish to ask themselves what *happens* to the magnetic field if we change our point of view to an inertial reference frame that is moving with the velocity of the charge. Hmmm, it seems as though in *this* frame there *is no magnetostatic field* where in the frame where the charge was moving, there was one! Yet somehow, the *physics* observed must be the same! This will take a bit of work, in some future E&M course, to arrange, and the theory that consistently permits it all to work out is the *theory of special relativity*.

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:

$$\vec{F}_{12} = -q_1 \vec{v}_1 \times k_m q_2 \left(\frac{\vec{v}_2 \times \hat{r}_{12}}{r_{12}^2} \right) = -\frac{k_m q_1 q_2}{r_{12}^2} (\vec{v}_1 \times (\vec{v}_2 \times \hat{r}_{12})) \quad (499)$$

Similarly, the force on q_2 due to q_1 is:

$$\vec{F}_{21} = q_2 \vec{v}_2 \times k_m q_1 \left(\frac{\vec{v}_1 \times \hat{r}_{12}}{r_{12}^2} \right) = \frac{k_m q_1 q_2}{r_{12}^2} (\vec{v}_2 \times (\vec{v}_1 \times \hat{r}_{12})) \quad (500)$$

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 78: 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; 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”!

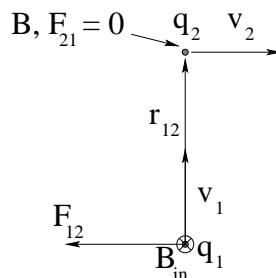


Figure 78: Newton's Third Law fails for this arrangement because the field (and hence force) at q_2 due to q_1 is *zero* while the field (and hence force) at q_1 due to q_2 is *not zero*!

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?

7.3: The Biot-Savart Law

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. As noted above, the magnetic fields of individual particles were largely beyond the experimental reach of eighteenth and nineteenth century physicists (which is where most of the electromagnetism we learn in this book was discovered). In fact, 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.
- They were the “artificial” fields generated by *current carrying wires* in the laboratory, under human-controlled conditions.

We will now focus on the second of these, current carrying wires, because historically Ampere and Biot-Savart used measurements of forces between current carrying wires to establish most of the important laws concerning the generation of magnetic fields, laws that in fact explain natural magnetism as well.

Consider a current-carrying wire and the microscopic model for conduction that we worked with in detail a few chapters back. In this model, a conductor can be viewed as an (electrically neutral) collection of fixed charge and mobile charge. The mobile charge carriers all have charge q (assumed positive, although of course q could be either sign and in general will actually be negatively charged electrons in metals). The density of carriers per unit volume is n . The wire has cross sectional area A , and we will initially assume that the wire is “thin” compared to the distance to the point of observation so we can avoid having to consider variations in the distance from a differential sized chunk of the wire to the point of observation \vec{r} . The geometry is shown in figure 79.

We know from the last section that the field of just *one* charge, say the one very close to the

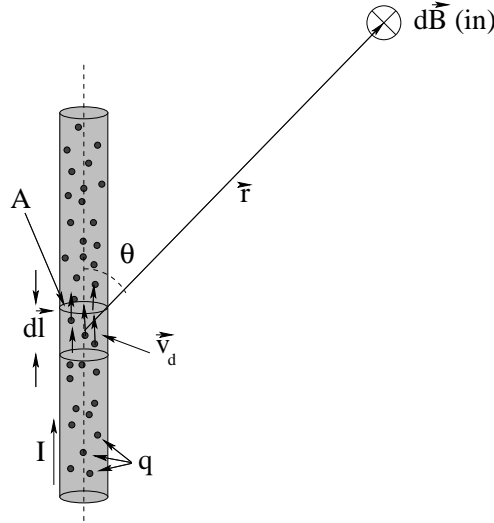


Figure 79: The microscopic model of conduction used to derive the Biot-Savart Law from the magnetic field of a point charge.

origin of \vec{r} , at the point \vec{r} is:

$$\vec{B} = \frac{\mu_0}{4\pi} \frac{q(\vec{v}_d \times \hat{r})}{r^2} \quad (501)$$

If we take a very small chunk of wire (and remember, we are exaggerating its width, which is really small relative to r) of length $d\ell$ then *all* of the charges in the volume $A d\ell$ are moving with velocity \vec{v}_d in the direction of the current, and have approximately the same vector \vec{r} to the point of observation. The total field of *all* of these charges, in any chunk that contains enough charges so that the idea of the density of charge carriers makes sense, is produced by the total charge in this volume.

That charge is:

$$dQ = nqAd\ell \quad (502)$$

and the field it produces is thus:

$$\begin{aligned} d\vec{B} &= \frac{\mu_0}{4\pi} \frac{dQ(\vec{v}_d \times \hat{r})}{r^2} \\ &= \frac{\mu_0}{4\pi} \frac{nqAd\ell(\vec{v}_d \times \hat{r})}{r^2} \end{aligned} \quad (503)$$

We now play a clever trick. We make a *vector* $d\vec{\ell} = d\ell\hat{v}_d$ that points in the same direction as \vec{v}_d , leaving behind its magnitude v_d , so that this equation becomes:

$$d\vec{B} = \frac{\mu_0}{4\pi} \frac{nqAv_d(d\vec{\ell} \times \hat{r})}{r^2} \quad (504)$$

and identify $I = nqAv_d$ from our previous conduction model! This then becomes:

$$d\vec{B} = k_m \frac{I(d\vec{\ell} \times \hat{r})}{r^2} \quad (505)$$

which is the *Biot-Savart* (Bee-oh Sah-vahr) Law!

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 80 becomes:

$$d\vec{B} = k_m \frac{I(d\vec{\ell} \times (\vec{r} - \vec{r}_0))}{|\vec{r} - \vec{r}_0|^3} \quad (506)$$

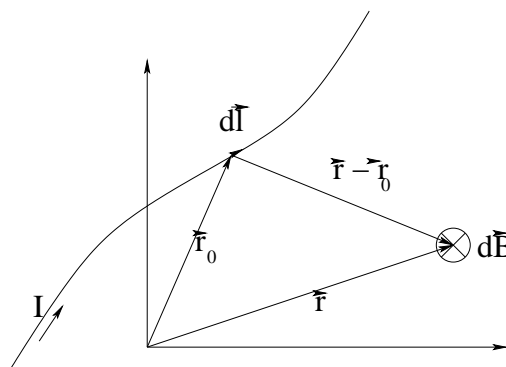


Figure 80: The Biot-Savart Law in general coordinates.

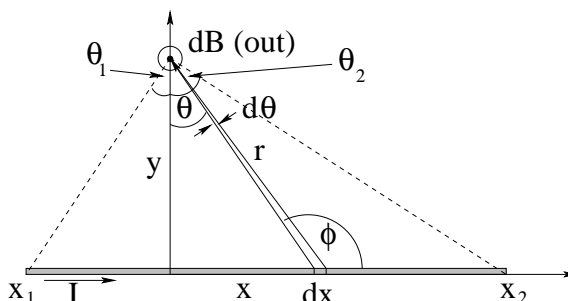
Wow! That looks a lot more complicated than our integral expressions for the electrostatic field! And so it is... we will have to work much harder to evaluate the magnetic field directly from a current (distribution), and the need to do twisty integrals of directed differential vectors as they are worked 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 straightforward 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, well, that is *close* to all of the examples one can do without being an integration god. 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).

For this course, we will cheerfully take the easy road and work through a nice set of relatively simple examples that is, as promised, most of or close to what you can do, period, before things get very complicated indeed.

7.4: Examples of Using the Biot-Savart Law to Find the Magnetic Field

Example 7.4.1: Magnetic Field of a Straight Wire Segment

Figure 81: 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 81 above, 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 for the field produced by this small differential chunk of the current as:

$$dB = \frac{k_m I \sin(\phi) dx}{r^2} \quad (507)$$

and hence (formally integrating both sides) we get:

$$B = \int_{x_1}^{x_2} \frac{k_m I \sin(\phi) dx}{r^2} \quad (508)$$

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:

$$\begin{aligned} x &= y \tan(\theta) \\ dx &= \frac{y d\theta}{\cos^2(\theta)} \end{aligned} \quad (509)$$

and

$$\sin(\phi) = \cos(\theta) \quad (510)$$

(think about it for a minute, it will make sense). The integral becomes:

$$B = \int_{\theta_1}^{\theta_2} \frac{k_m I y \cos(\theta) d\theta}{\cos^2(\theta) r^2} = \int_{\theta_1}^{\theta_2} \frac{k_m I y \cos(\theta) d\theta}{y^2} \quad (511)$$

(using $y = r \cos(\theta)$) and we finally obtain:

$$\begin{aligned} B &= \frac{k_m I}{y} \int_{\theta_1}^{\theta_2} \cos(\theta) d\theta \\ &= \frac{k_m I}{y} (\sin(\theta_2) - \sin(\theta_1)) \end{aligned} \quad (512)$$

(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 $\theta_1 \rightarrow -\pi/2$ and $\theta_2 \rightarrow \pi/2$, 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:

$$B_\infty = \frac{2k_m I}{y} \quad (513)$$

Again note the analogy with electric field, with $k_e \rightarrow k_m$ and $\lambda \rightarrow I$, 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.**

This latter is, as we already saw in our original discussions of the magnetic field, a useful and general rule as it will always work out to be the direction given by the cross product for a short segment of current.

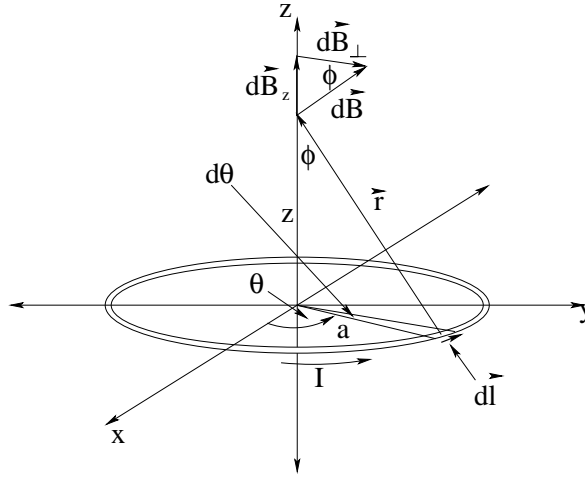
Example 7.4.2: Field of a Circular Loop on its Axis

Figure 82: The geometry and coordinates used to compute the magnetic field of a circular loop wire carrying a current I on its axis of symmetry.

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 $d\vec{l}$:

$$d\vec{B} = k_m \frac{I d\vec{l} \times \hat{r}}{r^2} \quad (514)$$

As you can see, the direction of this infinitesimal field element is in the plane formed by \vec{r} and the z -axis, perpendicular to \vec{r} in the right-handed direction. The magnitude of this field element is:

$$dB = k_m \frac{I dl}{r^2} \quad (515)$$

We can easily find the components of $d\vec{B}$ parallel and perpendicular to z using the angle ϕ :

$$dB_z = k_m \frac{I dl}{r^2} \sin(\phi) \quad (516)$$

$$dB_{\perp} = k_m \frac{I dl}{r^2} \cos(\phi) \quad (517)$$

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 $\cos(\phi) = z/r$:

$$dB_z = k_m \frac{I dl}{r^3} a \quad (518)$$

$$dB_{\perp} = k_m \frac{I dl}{r^3} z \quad (519)$$

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:

$$dB_z = k_m \frac{Idl a}{r^3} \quad (520)$$

$$dB_x = k_m \frac{Idl z}{r^3} \cos(\theta) \quad (521)$$

$$dB_y = k_m \frac{Idl z}{r^3} \sin(\theta) \quad (522)$$

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 $dl = a d\theta$, substitute this in, and we are ready to go:

$$\begin{aligned} B_z &= k_m \int_0^{2\pi} \frac{I a^2 d\theta}{r^3} \\ &= k_m \frac{I 2\pi a^2}{r^3} \end{aligned} \quad (523)$$

$$B_x = k_m \int_0^{2\pi} \frac{I a z}{r^3} \cos(\theta) d\theta = 0 \quad (524)$$

$$B_y = k_m \int_0^{2\pi} \frac{I a z}{r^3} \sin(\theta) d\theta = 0 \quad (525)$$

We conclude that:

$$\vec{B} = k_m \frac{I 2\pi a^2}{(a^2 + z^2)^{3/2}} \hat{z} \quad (526)$$

It is instructive to write this in terms of the *magnetic moment* of the loop, $\vec{m} = I\pi a^2 \hat{z}$:

$$\vec{B} = \frac{2k_m \vec{m}}{(a^2 + z^2)^{3/2}} \quad (527)$$

which is *exactly the same form* as that of the electric field on the axis of an electric dipole, $\vec{E} = 2k_e \vec{p}/(a^2 + z^2)^{3/2}$, that we derived several weeks ago, with the substitution of k_m for k_e and \vec{p} for \vec{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 $z \gg a$:

$$\vec{B} = \frac{2k_m \vec{m}}{z^3} \quad (528)$$

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 7.4.3: Field of a Revolving Ring of Charge on its Axis

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

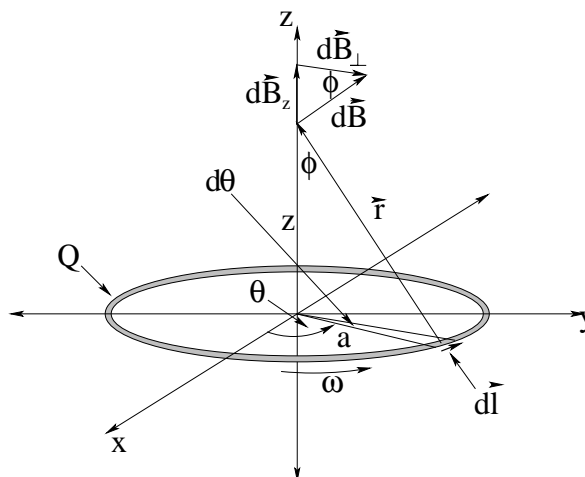


Figure 83: The geometry and coordinates used to compute the magnetic field on the axis of symmetry of a circular ring of charge Q revolving at angular velocity ω .

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:

$$I = \frac{Q}{T} = \frac{Q\omega}{2\pi} \quad (529)$$

The field is thus obtained by doing the exact same integrals as before:

$$\vec{B} = k_m \frac{I 2\pi a^2}{(a^2 + z^2)^{3/2}} \hat{z} = k_m \frac{Q\omega a^2}{(a^2 + z^2)^{3/2}} \hat{z} \quad (530)$$

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 $k_m 2\vec{m}/z^3$ 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*.

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. In a sense, this is conceptually one of the most difficult things to see, because it requires math that you almost certainly haven't seen yet. There is also little point in pursuing the formally correct "best" path – as we noted above, 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.

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 prize⁶⁶. 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 Stoke's Theorem Free discussion⁶⁷ by thinking once again about a single, infinitely long, straight line of current I . We recall from our Biot-Savart derivation above that:

$$B = \frac{2k_m I}{r} = \frac{\mu_0 I}{2\pi r} \quad (531)$$

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 84.

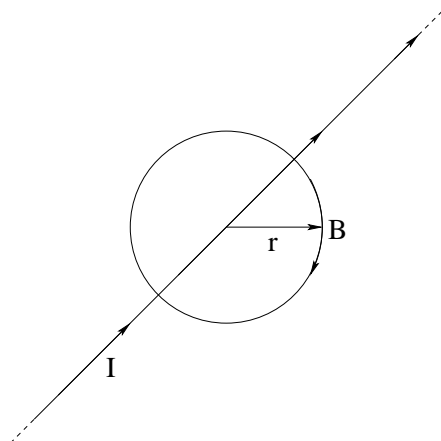


Figure 84: An infinitely long straight wire carries current I and has a magnetic field that goes *around* the wire in circular loops of constant magnitude in the right-handed sense.

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 multiplied a field that went down like $1/r^2$ by an area that went

⁶⁶A 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, you can't...

⁶⁷That, and the curl, are the math that we would be using if we were going to do this "right".

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:

$$B \times 2\pi r = \frac{\mu_0 I}{2\pi r} \times 2\pi r = \mu_0 I \quad (532)$$

So far, this is only suggestive. However, consider the geometry of figure 85. where the current I

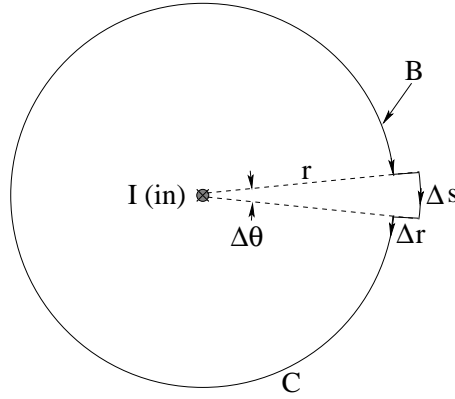


Figure 85: A circular path of radius r around a long straight wire with a “notch” of angular width $\Delta\theta$ and radius $r + \Delta r$.

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:

$$\begin{aligned} B(2\pi - \Delta\theta)r + B\Delta s &= B(2\pi - \Delta\theta)r + B(r + \Delta r)\Delta\theta \\ &= \frac{\mu_0 I}{2\pi r}(2\pi - \Delta\theta)r + \frac{\mu_0 I}{2\pi(r + \Delta r)}(r + \Delta r)\Delta\theta \\ &= \frac{\mu_0 I}{2\pi r}2\pi r - \frac{\mu_0 I}{2\pi}\Delta\theta + \frac{\mu_0 I}{2\pi}\Delta\theta \\ &= \mu_0 I \end{aligned} \quad (533)$$

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:

$$Bds = \frac{\mu_0 I}{2\pi r} r d\theta = \frac{\mu_0 I}{2\pi} d\theta \quad (534)$$

If we clean up the geometry of this, picking a path element along C with a *vector* length $d\vec{\ell}$ and selecting only the component parallel to \vec{B} at that point with the dot product, we get:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \oint_C \frac{\mu_0 I}{2\pi r} \hat{\theta} \cdot r d\theta \hat{\theta} = \frac{\mu_0 I}{2\pi} \int_0^{2\pi} d\theta = \mu_0 I \quad (535)$$

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{B} always cancelling the $1/r$ dependence of \vec{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"⁶⁸.

This leads us to write the previous equation in the following carefully selected form:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 I_{\text{through } C} = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \quad (536)$$

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!

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* ⁶⁹. This is the way I'd like you to practice writing Ampere's Law, although in *application* below we'll often just add up the total current through any given loop by inspection.

Also note that $\mu_0 = 4\pi k_m$ – a form that is rather reminiscent of the $4\pi k_e$ we saw in Gauss's Law for Electricity. Keep that loosely in mind for later, as eventually it will help us do a simple but enormously important piece of arithmetic without a calculator.

Another Maxwell Equation! 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) and in fact we usually derive the Biot-Savart Law from *it* in modern times, or better yet skip directly to relativistically correct, retarded descriptions of the electromagnetic field that are *really* difficult to evaluate (but correct).

For now, though, and in this course, we'll content ourselves with *the* handful of problems where Ampere's Law can be used to evaluate the magnetic field. We're basically going to do all of them.

⁶⁸This 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.

⁶⁹Hint: 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?

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 to 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).

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:

$$\begin{aligned}
 \oint_C \vec{B} \cdot d\vec{\ell} &= \mu_0 I_{\text{thru } C} \\
 B_t \oint_C d\ell &= \mu_0 I \\
 B_t 2\pi r &= \mu_0 I \\
 B_t &= \frac{\mu_0 I}{2\pi r}
 \end{aligned} \tag{537}$$

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

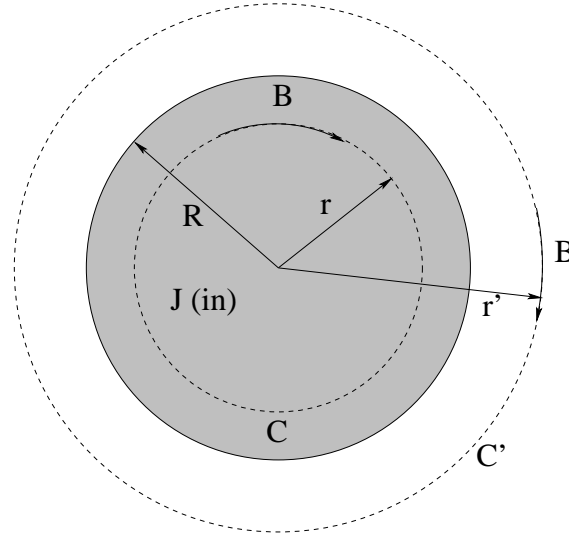


Figure 86: An infinitely long thick wire of circular radius R carries a current I into the page as drawn. We would like to find the magnetic field in all space.

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 86. 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{J} into the page:

$$\vec{J} = \frac{I}{\pi R^2} \hat{z} \quad (538)$$

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. 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 ($r < R$):

$$\begin{aligned} \oint_C \vec{B} \cdot d\vec{\ell} &= \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \\ B_t \oint_C d\ell &= \mu_0 J \int_{S/C} dA \\ B_t 2\pi r &= \mu_0 J \pi r^2 \\ B_t &= \frac{\mu_0 J \pi r^2}{2\pi r} = \frac{\mu_0 I r}{2\pi R^2} \end{aligned} \quad (539)$$

where we have selected a right-handed normal \hat{n} into the page so that the dot product of \vec{J} and \hat{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 ($r' > R$):

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA$$

$$B_t 2\pi r' = \mu_0 I$$

$$B_t = \frac{\mu_0 I}{2\pi r'} \quad (540)$$

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.

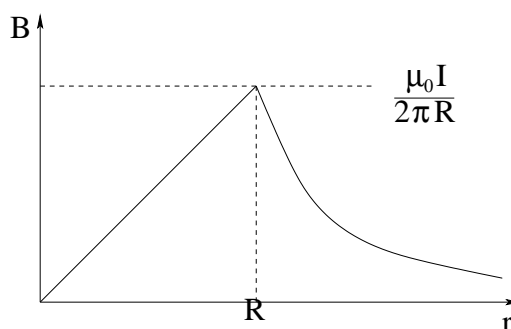


Figure 87: $B(r)$ for a long thick wire of radius R carrying a current I . Note that the field increases linearly inside of the wire and reaches a maximum value on the surface of the wire. Outside it drops off like $1/r$. Although the field is continuous, its derivative (slope) is not; it jumps at $r = R$.

We crudely plot the field as a function of r in figure 87. 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 88 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 enough for any snugly wrapped coil that is (much) longer than its diameter.

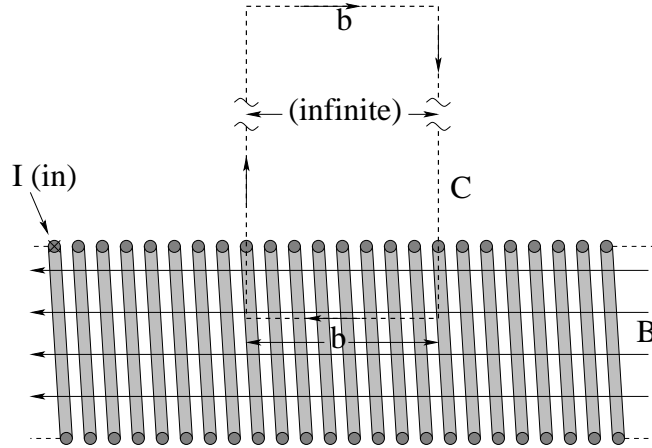


Figure 88: A cross-sectional view of an infinitely long solenoid with n turns per unit length, cross-sectional area A , carrying current I in each turn. The field both inside and outside of the solenoid is parallel to the axis of the solenoid (from symmetry), leading to the Amperian Path shown.

If you examine figure 88, 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 *parallel* 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 component 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”, wherever 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:

$$\begin{aligned}
 \oint_C \vec{B} \cdot d\vec{\ell} &= B_z b + 0(\text{left}) + 0(\text{top}) + 0(\text{right}) = \mu_0 I_{\text{thru } C} \\
 B_z b &= \mu_0 n b I \\
 B_z &= \mu_0 n I = \frac{\mu_0 N I}{L}
 \end{aligned} \tag{541}$$

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

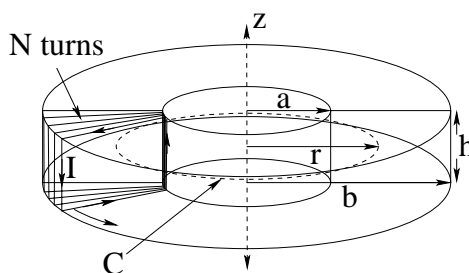


Figure 89: A cross-sectional view of a toroidal solenoid with N turns, and a rectangular cross-section with inner radius a , outer radius b , and height h , carrying current I in each turn. The field both inside the solenoid is concentric to the vertical axis of the torus (from symmetry and the right hand rule), leading to the Amperian Path shown.

In figure 89 above a toroidal⁷⁰ 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 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 closer 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:

$$\begin{aligned}\oint_C \vec{B} \cdot d\vec{\ell} &= \mu_0 I_{\text{thru } C} = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \\ B 2\pi r &= \mu_0 N I \\ B_t &= \frac{\mu_0 N I}{2\pi r}\end{aligned}\tag{542}$$

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

⁷⁰Wikipedia: <http://www.wikipedia.org/wiki/Torus>. A torus is a “doughnut shape”, usually with a circular cross section.

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.

Example 7.6.5: Infinite Sheet of Current

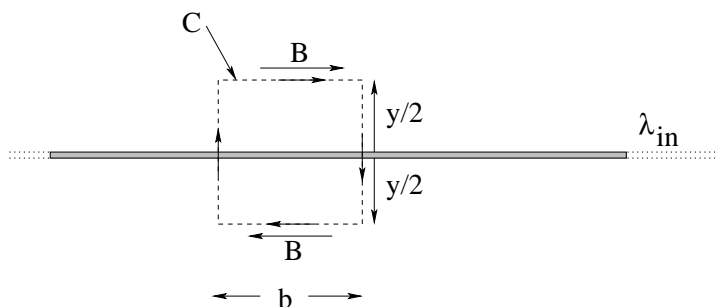


Figure 90: A side view of an infinite sheet of conductor carrying a current (per unit length) λ into the page. The field due to the sheet is symmetric up and below the sheet as drawn, and must point parallel to the sheet because every point is in the middle of the infinite plane (as usual). Any up-down 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 infinite solenoid, with sides perpendicular to the field.

In figure 90 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 :

$$\begin{aligned}\oint_C \vec{B} \cdot d\vec{\ell} &= \mu_0 I_{\text{thru } C} \\ 2B_{||}b &= \mu_0 \lambda b \\ B_{||} &= \frac{\mu_0 \lambda}{2}\end{aligned}\tag{543}$$

where $B_{||}$ is the magnitude of the component of \vec{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:

$$B_{||} = \mu_0 \lambda \tag{544}$$

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:

$$\lambda = nI \tag{545}$$

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:

$$B_{||} = \mu_0 \lambda = \mu_0 nI \tag{546}$$

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: Summary

Yes, this week is long enough, and has enough content, that it is worth a summary. 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} , although it is 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 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 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.

Finally, we obtained from the Biot-Savart Law a new equation we called *Ampere's Law* after its discoverer that is *consistent* with it (one can derive the Biot-Savart Law from Ampere's Law and, with some effort, vice versa) but that inherits its *flaw* that it is essentially a static result. We did find Ampere's Law to be remarkably useful for finding the *static* 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 still *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 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 electrostatics and magnetostatics* 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!*, 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 difficulty at all doing work on charged particles!

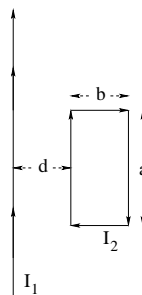
Homework for Week 7

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.



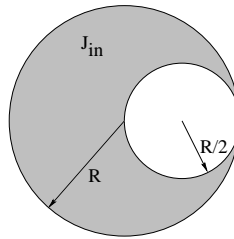
An infinitely long straight wire carries a current I_1 in the $+z$ direction. At $x = d$ there is a rectangular loop of current I_2 in the $x - z$ plane, with two sides of length a parallel to the long wire and two sides of length b perpendicular to the long wire. The current in the wire segment nearest the long wire is *parallel* to the current I_1 in the $+z$ direction. Find the net force acting on the rectangular loop.

Problem 3.

Using Ampere's Law, find the magnetic field in all space produced by:

- A solid conducting cylinder carrying a total current I .
- Two cylindrical conductings shells carrying opposite currents (each equal to I in magnitude). The inner one has radius a , the outer one b .
- A solenoid with N turns and length L carrying current I in each turn (inside only, far from the ends).
- A toroidal solenoid with N turns, inner radius a , outer radius b .
- An infinite plane sheet of current into the paper (above and below the sheet).

This more or less exhausts the *kinds* of possible problems where one can find the magnetic field using Ampere's Law. Most were examples in lecture, so this forces you to recapitulate on your own what you saw presented there.

Problem 4.

A cylindrical conductor of radius R aligned with the z direction has a cylindrical *hole* of radius $R/2$ centered at $x = R/2$ also aligned with the z direction. The conductor carries a *current density* $\vec{J} = J\hat{z}$ (and obviously $\vec{J} = 0$ in the hole). Find the magnetic field at all points inside the hole.

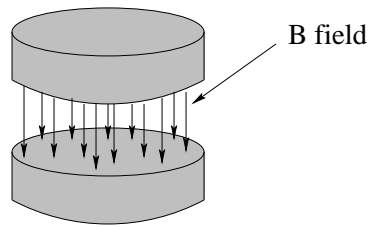
Problem 5.

Using the Biot-Savart law:

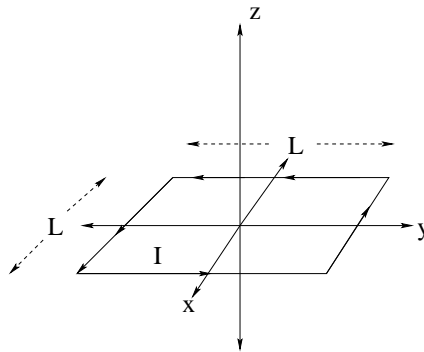
- Find the \vec{B} -field on the z axis of a circular current loop of radius a and N turns carrying a current I in the $x - y$ plane (centered on the origin).
- Set up the integral to be done to find the \vec{B} -field on the z axis of a *disk* in the $x - y$ plane of uniform charge density σ and radius a that is rotating with angular frequency ω around the z axis. (A) Do this integral (requires integration by parts a couple of times).

Problem 6.

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 that is rotating at angular velocity ω on a) 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 a ball spinning around the z axis. Note that it is quite a bit of work to actually derive this result (though it can be done). 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 given charge-current density to discover the (far) field “for free”.

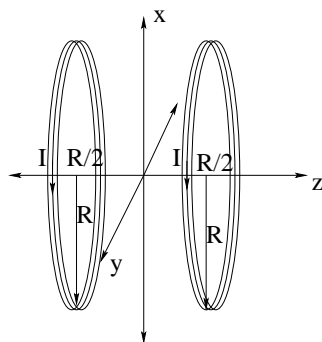
Problem 7.

Show that a uniform magnetic field that has no fringing field violates Ampere's law. Use a rectangular closed curve C that lies partly inside, and partly outside, the region of confined field. 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 8.

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 9.

(A) A pair of Helmholtz coils 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 that at $z = 0$: $\frac{dB_z}{dz} = 0$ and $\frac{d^2B_z}{dz^2} = 0$. This means that the magnetic field is quite “flat” in the middle of a Helmholtz coil.

Advanced Problem 10.

Find the magnetic field on the axis of a uniform disk of charge with radius R , mass M , and charge Q . This should be fairly easy to *set up* at this point and everybody ought to be able to do it. The resulting integral over r , however, will require integration by parts to solve, in particular

$$\int \frac{r^3 dr}{(r^2 + z^2)^{3/2}}$$

If you make $u = r^2$ and $v = (r^2 + z^2)$ this is pretty easy, but it is still a bit difficult for non-majors. A good challenge problem, though, for non-majors who want to improve their math and physics skills!

Express the answer in terms of the magnetic moment of the disk (computed previously) and show that its limiting form as $z \gg R$ is that of a dipole. Note also that this is the spinning disk that you demonstrated would *precess* in the previous chapter when placed in a strong field! At this point you should understand spinning disks of charge (as dipoles) pretty well!

V: Electrodynamics

Week 8: Faraday's Law and Induction

(Est 2/25-3/4)

- 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*:

$$\oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (547)$$

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:

$$\phi_i = \sum_{j \neq i} M_{ij} I_j + L_i I_i \quad (548)$$

where the M_{ij} are called the *mutual inductances* between the i th and j th loops and L_i is the *self inductance* of the i th 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:

$$B = \frac{\mu_0 N I}{L} \quad (549)$$

(parallel to the solenoid axis). The current I creates a flux *per turn* that is equal to:

$$\phi_t = BA = \frac{\mu_0 N A I}{L} \quad (550)$$

where A is the cross-sectional area of the solenoid. The total flux is thus:

$$\phi = N B A = \frac{\mu_0 N^2 A I}{L} = L_s I \quad (551)$$

where L_s is the self-inductance of the solenoid. Clearly:

$$L_s = \frac{\mu_0 N^2 A}{L} \quad (552)$$

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:

$$1 \text{ Henry} = \frac{1 \text{ Volt} - \text{Second}}{\text{Ampere}} = 1 \text{ Ohm} - \text{Second} \quad (553)$$

- In terms of inductance:

$$V_L = -L \frac{dI}{dt} \quad (554)$$

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:

$$\phi_t = \frac{\mu_0 N_1 A I_1}{L} \quad (555)$$

for the first solenoid, so:

$$\phi_2 = N_2 \frac{\mu_0 N_1 A I_1}{L} \quad (556)$$

is the total flux through the second solenoid due to the current in the first. Thus:

$$M_{21} = \frac{\phi_2}{I_1} = \frac{\mu_0 N_1 N_2 A}{L} = M_{12} = M \quad (557)$$

8.1: Magnetic Forces and Moving Conductors

Last week we saw that our study of the sources of the magnetic field, even before we reconsider the forces produced by those fields, are starting to raise red flags concerning the consistency of electromagnetic theory. As we begin this week, Newton's Third Law is toast, directly violated by magnetic forces between moving charged particles, and we ought to be very worried about things like momentum conservation that we derived from Newton's Third Law. 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 invariant currents. Finally, I noted that magnetic forces, by their very nature and defining equation, can do no work.

This week we will *begin* by considering certain puzzles associated with this last incontestable fact. Under some very easily constructible scenarios, it certainly *looks* like magnetic fields do work. However, when we take advantage of the freedom we *should* have in physics to *change inertial reference frames*, we can do some surprising things (like make magnetic fields and forces acting on moving charged particles vanish entirely). Changing frames ought not to alter the total force or classical motion produced by the total force, but this means that somehow *magnetic fields must be able to transform into electric fields* and vice versa as we change frames. Electric fields can indeed do work, so this might actually resolve our paradox!

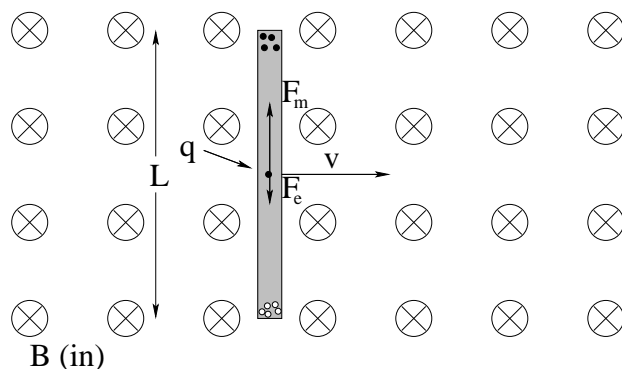


Figure 91: A conducting rod of length L moving through a uniform magnetic field into the page. The field *polarizes* the free charge in the rod until a region of crossed fields is produced.

To see the nature of the difficulty, we begin with a very simple picture – a conducting rod of length L moving through a uniform magnetic field at right angles to the field as show in figure 91. 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 on its sign.

The magnetic force on any given carrier is thus:

$$\vec{F}_m = q(\vec{v} \times \vec{B}) \quad (558)$$

which is *up* in 91. 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. 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!

We can easily deduce the condition for force balance for the charges in the rod proper:

$$\vec{F}_m + \vec{F}_e = 0 \quad (559)$$

or (since they are in opposite directions and the motion is at right angles to the magnetic field)

$$qvB = qE \quad (560)$$

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:

$$\Delta V = L \cdot E = (vL) \cdot B \quad (561)$$

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 charges even though we know that they cannot be doing any work. This leads us to a bit of a paradox – if the magnetic field isn't doing any work, what is?

To answer this, we note that we are examining what happens in a frame of reference in which the rod moves through a static magnetic field. Let's imagine that we have jumped *on* to the rod so that we are now at rest and the magnetic field is sweeping past us the *opposite way*. 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 the frame of reference we are in, and the magnetic field they are moving in isn't varying, it is constant in magnitude and direction! Yet things like the observed distribution of charge in this stationary frame has to *agree* with the distribution in the frame in which the rod moves, because 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 and generates the *same* internal electric field between the charges at the ends that we saw in the moving frame.

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 the charges can see 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 frame) the "induced" external field \vec{E}_{ind} must be equal in magnitude and opposite in direction to the polarization field:

$$E_{\text{ind}} = vB \quad (562)$$

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 by changing magnetic fields. 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

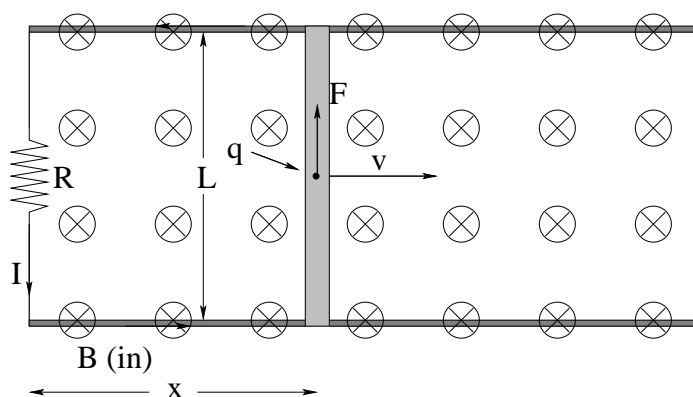


Figure 92: A conducting rod of mass M and length L moving through a uniform magnetic field into the page and sliding on *frictionless conducting rails* that are connected by a resistor R outside of the magnetic field. Current can flow around the loop thus formed.

In figure 92 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 it moved through the magnetic field of magnitude $E = vB$ that 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 potential difference induced across the rod as it moves is just:

$$\Delta V_{\text{ind}} = E_{\text{ind}}L = (vB)L \quad (563)$$

The last thing that has to trouble us is the sign of this potential difference. Again we need to appeal to physical invariance – in both frames we know that the magnetic force or induced electric force respectively must push the charges around the loop *counterclockwise* when v is to the right. Since we want Kirchoff's Loop Rule to be satisfied for this simple circuit loop and the voltage *decreases* across when we move across the resistor by IR , we expect the voltage around the loop to be positive, so that:

$$\Delta V_{\text{ind}} - IR = BLv - IR = 0 \quad (564)$$

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.

Where, exactly, is the field induced? What is it (in detail) inside of the conductor? This depends on the resistivity and current density associated with the entire conductive pathway, since we know that Ohm's Law is written as:

$$\vec{E} = \vec{J}\rho \quad (565)$$

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.

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 means that the induced electric field *forms a closed loop*. 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. *Electrostatic* fields begin and end on electric charges, but induced *electrodynamical* 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:

$$\oint_C \vec{F} \cdot d\vec{\ell} = 0 \quad (566)$$

for all closed loops C one can draw in space. The electrostatic field was conservative – if we let $\vec{F} = q\vec{E}$ and factored and cancelled q , we got:

$$\oint_C \vec{E} \cdot d\vec{\ell} = 0 \quad (567)$$

The *induced* *electrodynamical* field that appears in the loop, however, is *not conservative*! It has a nonzero integral around the loop:

$$\Delta V_{\text{ind}} = \oint_C \vec{E}_{\text{ind}} \cdot d\vec{\ell} = BLv \neq 0 \quad (568)$$

We recall that the whole point of a conservative field and its associated potential was that $\vec{E} = -\vec{\nabla}V$ (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{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):

$$|\Delta V_{\text{ind}}| = \oint_C \vec{E}_{\text{ind}} \cdot d\vec{\ell} = BLv = \frac{d(BLx)}{dt} \quad (569)$$

(because $v = \frac{dx}{dt}$) and, noting that $A = LX$ is the area inside of the loop we can write this as

$$|\Delta V_{\text{ind}}| = \oint_C \vec{E}_{\text{ind}} \cdot d\vec{\ell} = \frac{d(BA)}{dt} \quad (570)$$

which is just begging to be turned into the *flux of the magnetic field through the loop C*:

$$|\Delta V_{\text{ind}}| = \oint_C \vec{E}_{\text{ind}} \cdot d\vec{\ell} = \frac{d\phi_m}{dt} \quad (571)$$

where:

$$\phi_m = \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (572)$$

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 magic⁷¹ 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. 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.

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 this system, assuming only that it starts at time $t = 0$ moving at initial velocity v_0 to the right. $v(t)$, $I(t)$, and so on, find it all. Time to use Newton's Laws once again!

We begin with:

$$\begin{aligned} \Delta V_{\text{ind}} - IR &= 0 \\ \oint_C \vec{E}_{\text{ind}} \cdot d\vec{\ell} - IR &= 0 \\ BLv_x - IR &= 0 \end{aligned} \quad (573)$$

(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:

$$I = \frac{BLv_x}{R} \quad (574)$$

in the direction (counterclockwise) shown around the loop.

Next, compute the force acting on the rod. I flows *up* perpendicular to \vec{B} when v_x is positive, so Newton's Second Law becomes:

$$F_x = -ILB = m \frac{dv_x}{dt} \quad (575)$$

⁷¹You 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

$$\frac{dv_x}{dt} + \frac{B^2 L^2 v_x}{mR} = 0 \quad (576)$$

which is the usual first order, linear, homogeneous ordinary differential equation and is trivially integrable (see remarks in the math review section if the following doesn't make perfect sense to you):

$$\begin{aligned} \frac{dv_x}{dt} &= \left(-\frac{B^2 L^2}{mR} \right) v_x \\ \frac{dv_x}{v_x} &= \left(-\frac{B^2 L^2}{mR} \right) dt \\ \ln(v_x) = \int \frac{dv_x}{v_x} &= \int \left(-\frac{B^2 L^2}{mR} \right) dt \\ \ln(v_x) &= \left(-\frac{B^2 L^2}{mR} \right) t + C \\ v_x(t) &= v_0 \exp\left(-\frac{B^2 L^2 t}{mR} \right) \end{aligned} \quad (577)$$

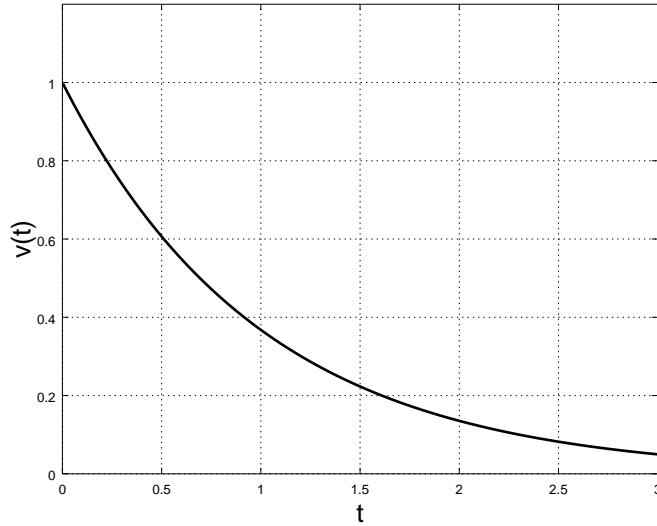


Figure 93: A plot of the exponential decay of the velocity of the rod as its initial kinetic energy is “burned” in heating the resistor with the induction-driven current that also slows it down. The units of the plot are v_0 (for v) and $\tau = \frac{mR}{B^2 L^2}$ (for t).

A plot of $v_x(t)$ is shown in figure 93 in units of v_0 and the exponential decay time $\tau = \frac{mR}{B^2 L^2}$.

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 (substituting in several results from above):

$$P_R = I^2 R = \frac{B^2 L^2 v^2}{R} = \frac{B^2 L^2 v_0^2}{R} \exp\left(-\frac{2B^2 L^2 t}{mR} \right) \quad (578)$$

The rate at which work is done on the rod is:

$$P_F = F \cdot v = -BLI \cdot v = -\frac{B^2 L^2 v_0^2}{R} \exp\left(-\frac{2B^2 L^2 t}{mR}\right) \quad (579)$$

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:

$$P_R + P_F = 0 \quad (580)$$

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 (we could easily find $x(t)$, for example, by integrating $\int v(t)dt$) and it all works out perfectly consistently. 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. Since 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 this electric field 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.

In the next section we will clearly state the conclusions of the first two chapters 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 the natural direction to specify the “positive” direction for the loop (clockwise as drawn in figure 92), then an *increasing* loop area and *increasing* flux produced a *negative* directed electric field (counterclockwise as drawn) and induced current that went the other way. This in turn made the force on the rod *negative* 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:

$$V_{\text{induced in } C} = \oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (581)$$

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{B} through the loop, then if the flux through the loop is increasing the E -field circulates the loop C in the negative (right handed) direction; if the flux through the loop is decreasing the E -field circulates around C in the positive direction.

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, note well that Faraday's Law states that an electric field will be induced around *arbitrary* loops C , not just loops C that correspond to the position in space of conductors! This is actually consistent with our reasoning in the very first section; we concluded that for the isolated (no conducting loop) rod moving in the magnetic field, it experienced an *external electric field* from the magnetic field sweeping over it in the frame where the rod is at rest and the field moves in the opposite direction. In fact, even in this problem where there is no loop at all the area swept out by the rod is $dA = Lv \, dt$

$$\Delta V_{\text{ind}} = -\frac{dBdA}{dt} = -BLv = -EL \quad (582)$$

so that the induced electric field is $E_{\text{ind}} = -Bv$ (where the minus sign means that the field points in the opposite direction to the “crossed fields” electric field that develops to cancel it).

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 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 author⁷², 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 it is 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.

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

⁷²Yes, 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...

magnetic flux through some surface S bounded by a closed curve C :

$$\phi_m(t) = \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (583)$$

you will soon realize that the flux ϕ_m can vary in time for any or all of *four reasons*:

- C can change in time (and hence so can S).
- The magnitude of \vec{B} can change in time.
- The angle between \vec{B} and \hat{n} can change in time because the direction of \vec{B} changes.
- The angle between \vec{B} and \hat{n} can change in time because the direction of \hat{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 makes 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.

0.0.1 Lenz's Law for changing C

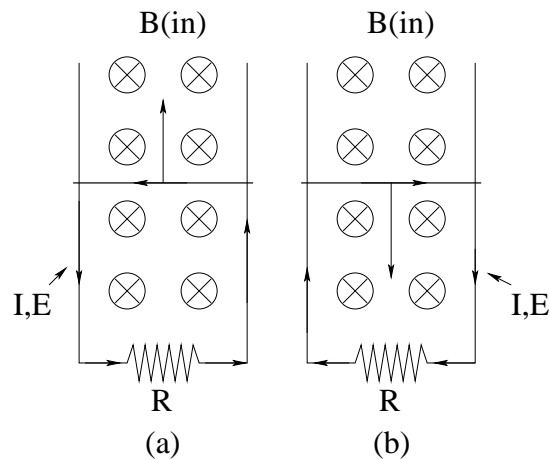


Figure 94: Illustration of \vec{E} -field direction for loops that change size. In (a) the loop is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise \vec{E} field and current opposes the existing field through the loop. In (b) the loop is getting 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.

We've already seen an example of this in our single meaningful example this far. If a plane loop C in a fixed magnetic field is *increasing in size*, then the induced field points in the opposite direction 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 94), 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.

0.0.2 Lenz's Law for changing B (magnitude)

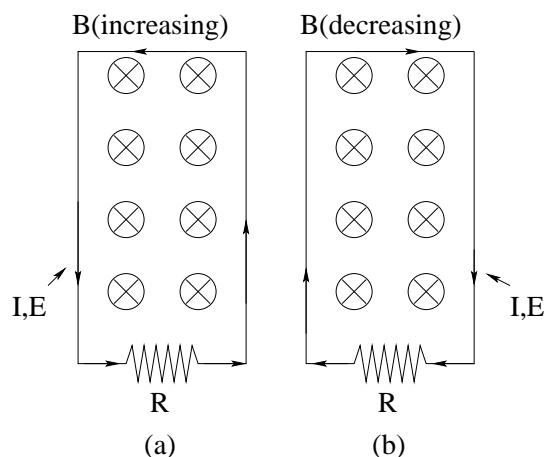


Figure 95: Illustration of \vec{E} -field direction when the magnitude of B through the loop changes. In (a) B is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise \vec{E} field and current opposes the existing field through the loop. In (b) 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 95 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* decreasing 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.

0.0.3 Lenz's Law for changing \vec{B} or \hat{n} direction

Now we imagine the shape of the loop C doesn't change, the magnetic field is constant in magnitude, but the loop's *orientation* in the magnetic field could be changing *or* the *direction* of the magnetic field could be changing. Note that both 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 \hat{n} are rotating into alignment about the dashed line axis shown (decreasing θ and hence increasing $\cos(\theta)$ and the flux) as shown in (a) of figure 96, the field direction and induced current are *clockwise* when viewed from above the loop to make the induced magnetic moment opposite to

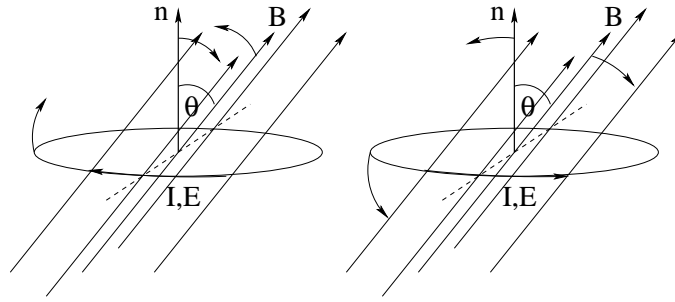


Figure 96: Illustration of \vec{E} -field direction when the direction of \vec{B} or the direction of the normal to the loop \hat{n} changes. In (a) $\cos(\theta)$ is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise \vec{E} field and current 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.

\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.

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*.

That is precisely the right direction for energy conservation to always hold for the system. We can breathe a sigh of relief!

Example 8.4.1: Wire and Rectangular Loop – Direction Only

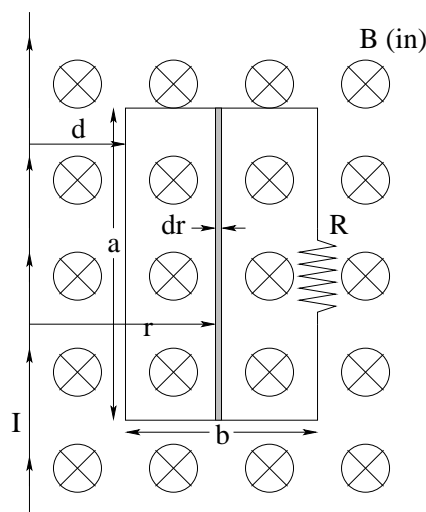


Figure 97: A long straight wire sits next to a rectangular loop of wire and carries a current I up as shown. The current in the long straight wire can be increased or decreased.

In figure 97 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.

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 for a current up – 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 the flux only changes *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 no current at all.

The 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.

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 distance 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.

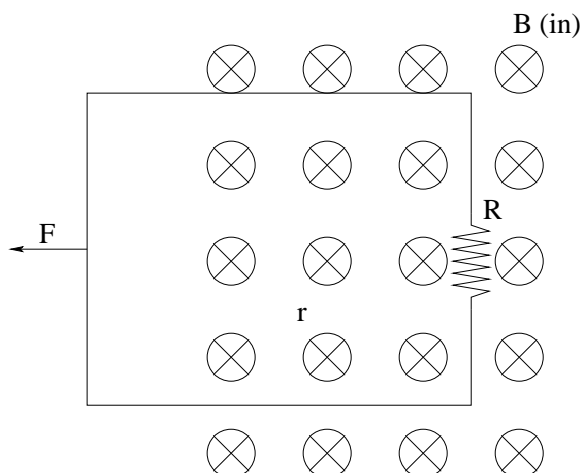
Example 8.4.2: Rectangular Loop Pulled from Field

Figure 98: A rectangular loop of wire is pulled out of a region of uniform magnetic field as shown.

In figure 98 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 force on this current (check) *resists the motion of the loop out of the field*.
- 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**

In figure 99 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

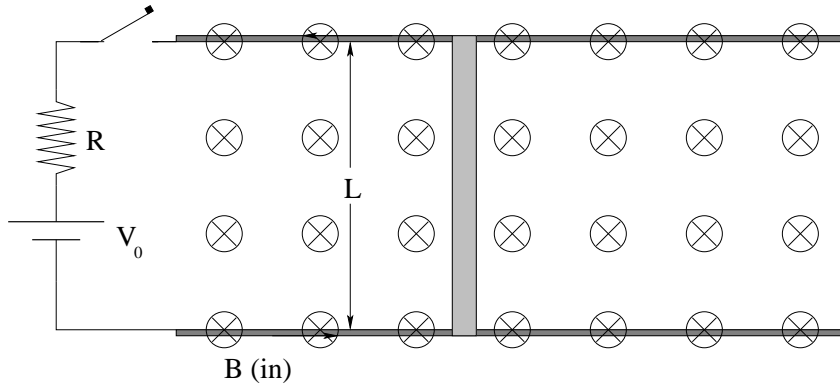


Figure 99: A conducting rod sits on conducting, frictionless rails and a switch is closed at $t = 0$ to send current through the loop thus formed. A magnetic field (into the page) exerts a force on the rod.

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:

- The voltage in the loop as a function of v , the (eventual) velocity of the rod.
- The current in the loop as a function of this voltage.
- The force on the rod as a function of this current.
- The *terminal* velocity of the rod, after the switch has been closed for a long time.
- The equation of motion of the rod as determined by the force.
- The velocity of the rod as a function of time.

This list lays out a very nice solution strategy. Using Faraday's Law

$$V_{\text{ind}} = -\frac{d\phi_m}{dt} = -\frac{dBLx}{dt} = -BLv \quad (584)$$

(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:

$$V_0 - BLv - IR = 0 \quad (585)$$

We can then solve for the current in the loop:

$$I = \frac{V_0 - BLv}{R} \quad (586)$$

and will circulate *clockwise* in the loop initially when v is small.

This lets us easily compute the force on the loop:

$$F = BLI = \frac{BLV_0 - B^2L^2v}{R} \quad (587)$$

and the terminal velocity, which we determine from the observation that the net force on the loop (and hence current in the loop) must be zero at the terminal velocity:

$$v_{\text{terminal}} = \frac{V_0}{BL} \quad (588)$$

Using the force equation we can easily write Newton's second Law and turn it into an equation of motion:

$$F = \frac{BLV_0 - B^2L^2v}{R} = Ma = M \frac{dv}{dt} \quad (589)$$

which we can rearrange into a *first order, linear, homogeneous, ordinary differential equation*:

$$\begin{aligned} \frac{dv}{dt} &= \frac{BLV_0 - B^2L^2v}{MR} \\ &= -\frac{B^2L^2}{MR} \left(v - \frac{V_0}{BL} \right) \\ \frac{dv}{\left(v - \frac{V_0}{BL} \right)} &= -\frac{B^2L^2}{MR} dt \\ \int \frac{dv}{\left(v - \frac{V_0}{BL} \right)} &= -\int \frac{B^2L^2}{MR} dt \\ \ln \left(v - \frac{V_0}{BL} \right) &= -\frac{B^2L^2}{MR} t + C \\ v - \frac{V_0}{BL} &= e^{-\frac{B^2L^2}{MR} t} * e^C \\ v(t) &= \frac{V_0}{BL} \left(1 - e^{-\frac{B^2L^2}{MR} t} \right) \end{aligned} \quad (590)$$

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*.

With this in hand we can easily integrate over time again to get $x(t)$, or differentiate it to get $a(t)$. 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.

8.6: Inductance

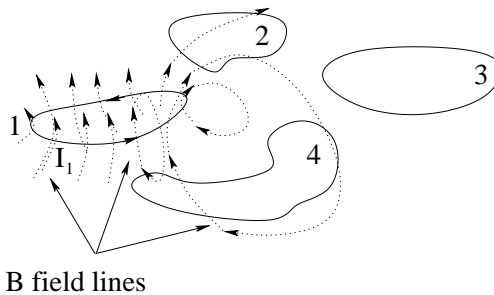


Figure 100: A set of current loops indexed by $i = 1, 2, 3, \dots$, fixed in space and carrying currents I_i . The B -field produced by (say) current I_1 swirls around the current and passes through both loop 1 and the other loops in the figure, creating both *self inductance* and *mutual inductance*.

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 100. 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 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 i th loop caused by the current in the j th loop is:

$$\begin{aligned}
 \phi_{ij} &= \int_{S_i/C_i} \vec{B}_j \cdot \hat{n}_i dA_i \\
 &= \frac{\mu_0}{4\pi} \int_{S_i/C_i} \left(\int_{C_j} \frac{I_j d\vec{l}_j \times (\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|^3} \right) \cdot \hat{n}_i dA_i \\
 &= \frac{\mu_0}{4\pi} \left(\int_{S_i/C_i} \int_{C_j} \frac{d\vec{l}_j \times (\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|^3} \cdot \hat{n}_i dA_i \right) I_j \\
 &= M_{ij} I_j
 \end{aligned} \tag{591}$$

where I've take some pains to label the coordinates with the object: \hat{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{r}_i the vector coordinate of a point on its surface; coordinates $d\vec{l}_j$ and \vec{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 know⁷³. The result is (eventually):

$$\begin{aligned}
 M_{ij} &= \frac{\mu_0}{4\pi} \left(\int_{S_i/C_i} \int_{C_j} \frac{d\vec{l}_j \times (\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|^3} \cdot \hat{n}_i dA_i \right) \\
 &= \frac{\mu_0}{4\pi} \oint_{C_1} \oint_{C_2} \frac{d\vec{l}_i \cdot d\vec{l}_j}{|\vec{r}_i - \vec{r}_j|}
 \end{aligned} \tag{592}$$

which is obviously symmetric under interchange of i and j :

$$M_{ij} = M_{ji} \tag{593}$$

for any two loops C_i and C_j carrying currents I_i and I_j respectively.

⁷³Wikipedia: 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.

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 i th and j th 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:

$$\begin{aligned}
 \phi_{ii} &= \int_{S_i/C_i} \vec{B}_i \cdot \hat{n}_i dA_i \\
 &= \frac{\mu_0}{4\pi} \int_{S_i/C_i} \left(\int_{C_i} \frac{\mu_0 I_i d\vec{l}_i' \times (\vec{r}_i - \vec{r}_i')}{|\vec{r}_i - \vec{r}_i'|^3} \right) \cdot \hat{n}_i dA_i \\
 &= \frac{\mu_0}{4\pi} \oint_{C_1} \oint_{C_1} \frac{d\vec{l}_i \cdot d\vec{l}_i'}{|\vec{r}_i - \vec{r}_i'|} I_i \\
 &= M_{ii} I_i \\
 &= L_i I_i
 \end{aligned} \tag{594}$$

where we define the *self-inductance* of the i th 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 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 i th circuit loop due to the currents in *all* of the loops:

$$\phi_i = L_i I_i + \sum_{j \neq i} M_{ij} I_j \tag{595}$$

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 i th loop:

$$V_i = -L_i \frac{dI_i}{dt} + \sum_{j \neq i} M_{ij} \frac{dI_j}{dt} \tag{596}$$

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:

$$V_L = -L \frac{dI}{dt} \tag{597}$$

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-Second²/Coulomb, or (since a Volt is a Joule/Coulomb) a Joule/Ampere².

Henries can, of course, also be expressed in terms of Webers – you *do* remember 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

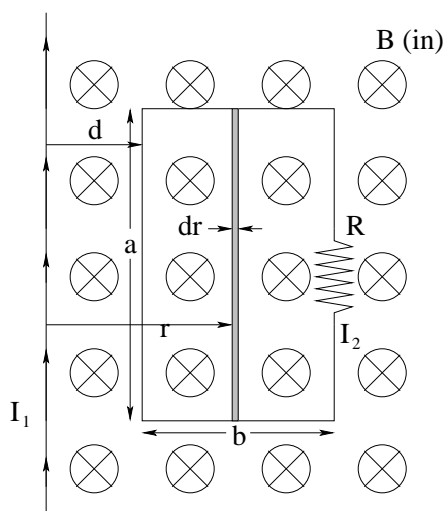


Figure 101: A long straight wire carrying a time-varying current $I_i(t)$ near a rectangular current loop induces a voltage V_2 in that loop, which in turn creates a current I_2 in that wire and a force \vec{F}_2 on the wire loop.

In figure 101 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 $I_1(t) = \frac{I}{T}t$, what is $\vec{F}_2(t)$? 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!):

$$B_1 = \frac{\mu_0 I_1}{2\pi r} \quad (598)$$

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:

$$d\phi_2 = \left(\frac{\mu_0 I_1}{2\pi r} \right) a dr \quad (599)$$

and

$$\begin{aligned} \phi_2 &= \left(\frac{\mu_0 I_1 a}{2\pi} \right) \int_d^{d+b} \frac{dr}{r} \\ &= \left(\frac{\mu_0 I_1 a}{2\pi} \right) \ln \left(\frac{d+b}{d} \right) \end{aligned} \quad (600)$$

- We can find the mutual inductance by dividing the flux by I_1 :

$$M_{21} = \frac{\phi_2}{I_1} = \left(\frac{\mu_0 a}{2\pi} \right) \ln \left(\frac{d+b}{d} \right) \quad (601)$$

(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!)

$$V_2 = -\frac{d\phi_2}{dt} = -\left(\frac{\mu_0 a}{2\pi} \right) \ln \left(\frac{d+b}{d} \right) \frac{dI_1}{dt} \quad (602)$$

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!)

$$V_2 - I_2 R = 0 \quad (603)$$

or

$$I_2(t) = \left(\frac{\mu_0 a}{2\pi R} \right) \ln \left(\frac{d+b}{d} \right) \frac{dI_1}{dt} \quad (604)$$

(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 *again* 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*⁷⁴

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: <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

In figure 102 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):

$$\begin{aligned} \oint_C \vec{B} \cdot d\vec{l} &= \mu_0 I_{\text{thru } C} \\ Bb &= \mu_0 \frac{N}{\ell} Ib \\ B &= \mu_0 \frac{N}{\ell} I \end{aligned} \tag{605}$$

⁷⁴Wikipedia: 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...?

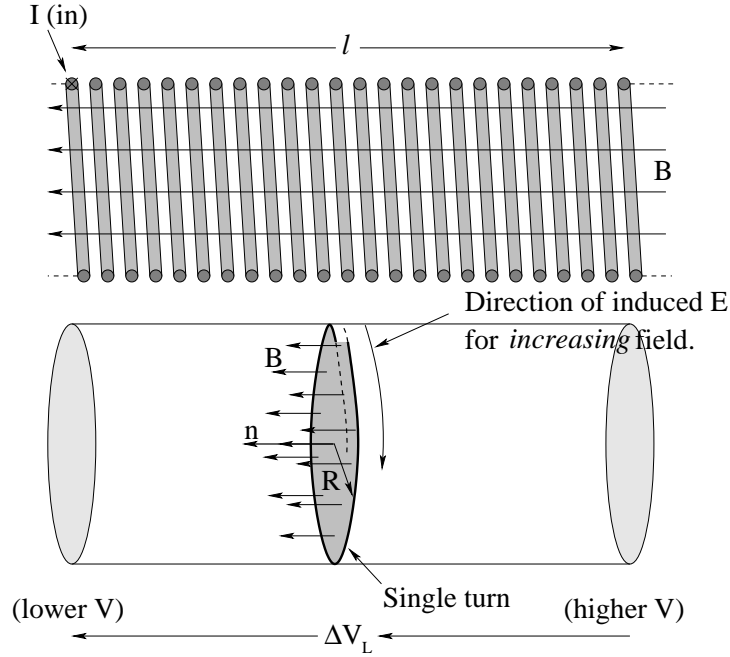


Figure 102: An ordinary (ideal) solenoid with N turns each carrying a current $I(t)$ is drawn above. The total flux through the solenoid is N times the flux through a single turn.

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 idealized 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:

$$\begin{aligned}\phi_{\text{turn}} &= \int_S \vec{B} \cdot \hat{n} dA \\ &= B\pi R^2 \\ &= \mu_0 \frac{N}{\ell} I \pi R^2\end{aligned}\tag{606}$$

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:

$$\phi_{\text{total}} = \left(\frac{\mu_0 N^2 \pi R^2}{\ell} \right) I\tag{607}$$

Finally, we find the self-inductance by noting that $\phi_{\text{total}} = LI$ so that:

$$L = \frac{\mu_0 N^2 \pi R^2}{\ell}\tag{608}$$

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:

$$\Delta V_L = -L \frac{dI}{dt} = I_0 (\omega L) \cos(\omega t) \quad (609)$$

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

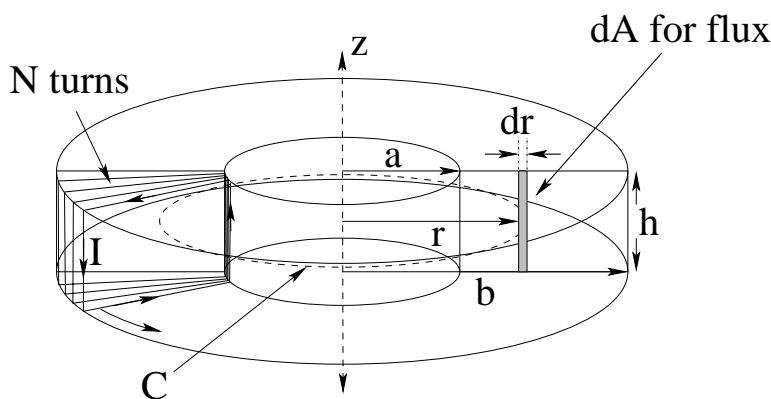


Figure 103: A tightly-wrapped toroidal solenoid with N turns produces a magnetic field inside that varies with r , but is approximately constant everywhere in a narrow strip of height h and width dr . The field is, of course, in the direction determined by the right hand rule, meaning that it points *in* to the page through the shaded strip we need to use to find the flux.

In figure 103 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:

- 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!
- 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

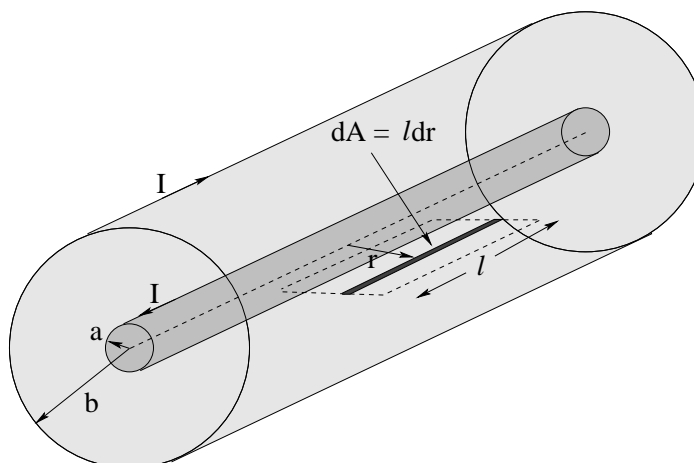


Figure 104: Coaxial cables have a self-inductance measured *per unit length*. At high frequencies the inductance only depends on the outer radius of the inner conductor a and the inner radius of the outer conductor b . A strip of area $dA = \ell dr$ is shown that may be of use in computing L/ℓ , the 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 104 a current $I(t)$ flows e.g. up the (long-straight) inner conductor and back on the outer one or vice versa. From Ampere's Law you can easily find the magnetic field in between the inner and outer conductors (where it is confined; at high frequencies all of the current will be on the surfaces and we can ignore current density and magnetic field inside the conductors themselves).

With the 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 is the flux per length ℓ 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, 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 archotypical 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.

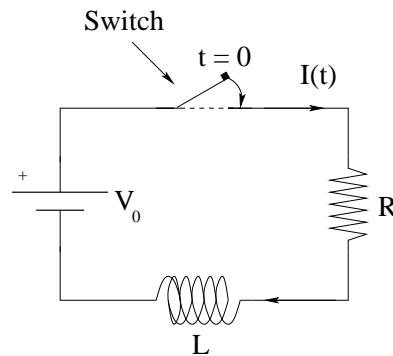


Figure 105: The archotypical direct current LR circuit. We generally assume the switch closes at time $t = 0$ with the current in the circuit $I(0) = 0$.

Let us, then, figure out a simple DC LR circuit, given in figure 105: 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. Henrys are Volt-Seconds/Ampere. We want a time constant in seconds, so it looks like:

$$\tau = \frac{L}{R} \quad (610)$$

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:

$$I(t) = \frac{V_0}{R} \left(1 - e^{-\frac{R}{L}t}\right) = \frac{V_0}{R} \left(1 - e^{-\frac{R}{L}t}\right) \quad (611)$$

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:

$$V_0 - IR - L \frac{dI}{dt} = 0 \quad (612)$$

We rearrange this to put it in the standard form of a first order, linear, inhomogeneous ordinary differential equation:

$$\frac{dI}{dt} + \frac{R}{L}I = \frac{V_0}{L} \quad (613)$$

At this point I shouldn't have to help you. We've now solved this equation several times over two semesters⁷⁵ – 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.

So here we go:

$$\begin{aligned} \frac{dI}{dt} + \frac{R}{L}I &= \frac{V_0}{L} \\ \frac{dI}{dt} &= \frac{V_0}{L} - \frac{R}{L}I \\ \frac{dI}{dt} &= -\frac{R}{L} \left(I - \frac{V_0}{R}\right) \\ \frac{dI}{\left(I - \frac{V_0}{R}\right)} &= -\frac{R}{L}dt \\ \int \frac{dI}{\left(I - \frac{V_0}{R}\right)} &= \int \left(-\frac{R}{L}\right) dt \\ \ln \left(I - \frac{V_0}{R}\right) &= -\left(\frac{R}{L}\right)t + C \\ \exp \left\{ \ln \left(I - \frac{V_0}{R}\right) \right\} &= \exp \left\{ -\left(\frac{R}{L}\right)t + C \right\} \\ I - \frac{V_0}{R} &= e^{-\left(\frac{R}{L}\right)t} e^C \\ I &= \frac{V_0}{R} + Ae^{-\left(\frac{R}{L}\right)t} \\ I(t) &= \frac{V_0}{R} \left(1 - e^{-\left(\frac{R}{L}\right)t}\right) \end{aligned} \quad (614)$$

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).

⁷⁵Approach 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.

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):

$$P_V = V_0 I(t) = \frac{V_0^2}{R} \left(1 - e^{-\left(\frac{R}{L}\right)t}\right) \quad (615)$$

Wow, that was easy!

The power burned in the resistor (negative – remember, this is energy that is all turned into (joule) heat(ing):

$$\begin{aligned} P_R &= V_R I(t) = (-I(t)R)I(t) = -I(t)^2 R \\ &= -\frac{V_0^2}{R} \left(1 - e^{-\left(\frac{R}{L}\right)t}\right)^2 \\ &= -\frac{V_0^2}{R} \left(1 - 2e^{-\left(\frac{R}{L}\right)t} + e^{-2\left(\frac{R}{L}\right)t}\right) \end{aligned} \quad (616)$$

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...):

$$\begin{aligned} P_L &= V_L I(t) = \left(-L \frac{dI}{dt}\right) I(t) \\ &= -\left\{L \frac{V_0}{R} \left(\frac{R}{L}\right) e^{-\left(\frac{R}{L}\right)t}\right\} \frac{V_0}{R} \left(1 - e^{-\left(\frac{R}{L}\right)t}\right) \\ &= -\frac{V_0^2}{R} \left(e^{-\left(\frac{R}{L}\right)t} - e^{-2\left(\frac{R}{L}\right)t}\right) \end{aligned} \quad (617)$$

Note that we used the fact that

$$V_L(t) = -L \frac{dI}{dt} = -V_0 e^{-\left(\frac{R}{L}\right)t} \quad (618)$$

is the voltage drop across the inductor just as:

$$V_R(t) = -IR = -V_0 \left(1 - e^{-\left(\frac{R}{L}\right)t}\right) \quad (619)$$

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 $I(t)$:

$$\begin{aligned} V_0 - IR - L \frac{dI}{dt} &= 0 \\ (V_0 - IR - L \frac{dI}{dt}) I(t) &= 0 \\ V_0 I(t) - I(t)^2 R - L \frac{dI}{dt} I(t) &= 0 \\ P_V + P_R + P_L &= 0 \end{aligned} \quad (620)$$

(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 somehow being *stored* in the inductor in the magnetic field. Then:

$$P_L = \frac{dU_L}{dt} = -LI \frac{dI}{dt} \quad (621)$$

or (multiplying by dt):

$$\begin{aligned} dU_L &= -LI dI \\ \int_0^{U_{\text{tot}}} dU_L &= \int_0^{I_0} -LI dI \\ U_{\text{tot}} &= \frac{1}{2}LI_0^2 \end{aligned} \quad (622)$$

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 us 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:

$$B = \frac{\mu_0 N I_0}{\ell} \quad (623)$$

and that vanishes at the ends of the solenoid (neglecting fringing fields). We showed above that the self-inductance of this ideal solenoid is:

$$L = \frac{\mu_0 N^2 A}{\ell} \quad (624)$$

Let's do an algebra-morph of the energy stored on the inductor:

$$\begin{aligned} U &= \frac{1}{2}LI^2 \\ &= \frac{\mu_0 N^2 A}{2\ell} I^2 \\ &= \frac{\mu_0^2 N^2 A \ell}{2\ell^2 \mu_0} I^2 \\ &= \frac{1}{2\mu_0} \frac{\mu_0^2 N^2 I^2}{\ell^2} A \ell \\ \Delta U &= \frac{B^2}{2\mu_0} \Delta \mathcal{V} \\ \frac{\Delta U}{\Delta \mathcal{V}} &= \frac{B^2}{2\mu_0} \end{aligned} \quad (625)$$

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*⁷⁶:

$$\eta_m = \frac{dU_m}{d\mathcal{V}} = \frac{B^2}{2\mu_0} \quad (626)$$

which strangely matches our similar equation (deduced from very similar considerations for the energy density in the *electric* field:

$$\eta_e = \frac{dU_e}{d\mathcal{V}} = \frac{1}{2}\epsilon_0 E^2 \quad (627)$$

There is something really sort of spooky about this – it is redolent⁷⁷ 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.

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:

$$\frac{dU}{d\mathcal{V}} = \frac{B(r)^2}{2\mu_0} \quad (628)$$

if we multiply this out:

$$dU = \frac{B(r)^2}{2\mu_0} d\mathcal{V} \quad (629)$$

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:

$$U = \frac{1}{2}LI^2 \quad (630)$$

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

$$d\mathcal{V} = 2\pi r h dr \quad (631)$$

Give it a shot, for homework. You can do it!

8.10: Eddy Currents

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”.

⁷⁶Physicsspeak for “inspired guess”...

⁷⁷Politespeak for “it stinks”...

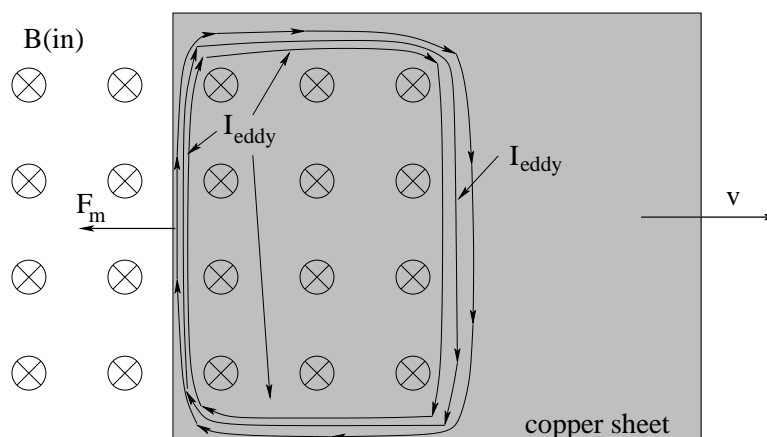


Figure 106: A sheet of copper being pulled rapidly out of a field has induced *eddy currents*. The forces from these currents, according to Lenz's Law, *resist* the motion, causing a magnetic “drag force” similar to that observed in the rod on rails problem. The kinetic energy of the object is transformed into heat by these currents (resistive Joule heating).

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 rapidly a varying magnetic field to directly heat metal pots placed in the field⁷⁸. 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 106 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.

⁷⁸Wikipedia: http://www.wikipedia.org/wiki/Induction_Cooker. This is actually a lovely article, and will introduce you to the notion of *skin depth*, as induction stovetops only tend to work on ferromagnetic pans (such as cast iron) because they have a small skin depth.

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* 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 phenomenon.

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 107 an external magnetic field into the page is applied to a (highly

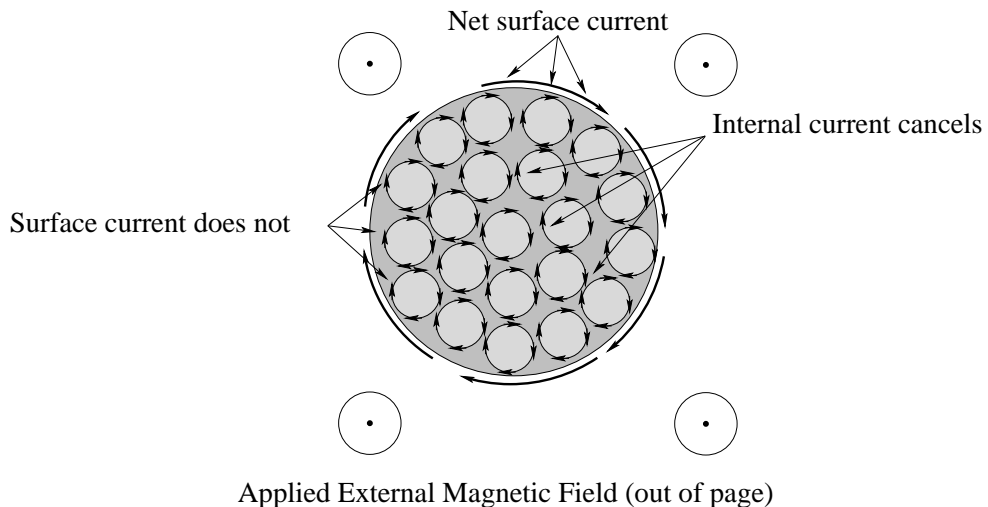


Figure 107: Wherever "atomic" magnetic current loops adjoin one another, the average current is zero. On the surface, however, there are no neighboring atoms, and the current loops there are not cancelled. They add (on average) into a *continuous surface current* not unlike that of a solenoid, so that the field everywhere in the interior is *reduced*.

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 material⁷⁹.

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.

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 Meissner Effect: [http://www.wikipedia.org/wiki/Meissner Effect](http://www.wikipedia.org/wiki/Meissner_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.

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 moments⁸⁰.

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

⁷⁹This 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.

⁸⁰Wikipedia: 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.

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 108.

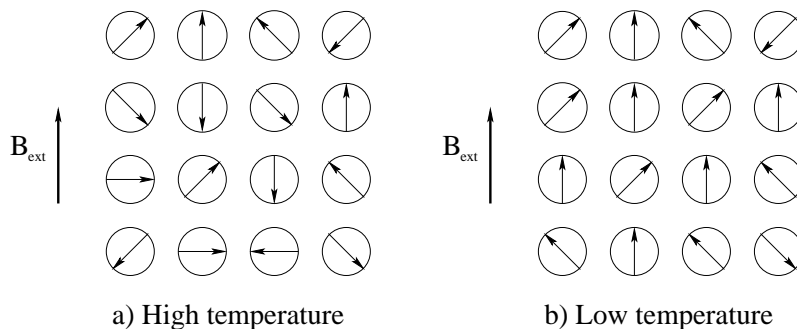


Figure 108: A lattice of “spins” at high temperature (a) and low temperature (b) is portrayed as a two dimensional cartoon. The direction of the arrows can be thought of as the directions of the angular momentum and hence magnetic moment of each atom, in a side view that reveals their rough degree of 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!*

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 109.

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

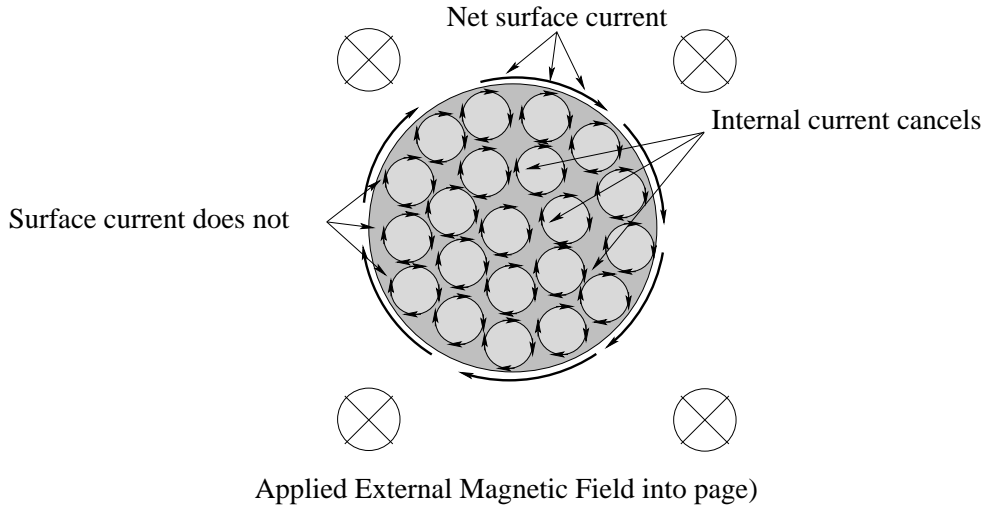


Figure 109: Just as was the case for a diamagnet, the internal currents of aligned magnetic moments cancel (on average) in the bulk of the material, but the surface currents *add*. The surface currents behave like the wires of a solenoid or sheet of current wrapped around the object to *increase* the total field inside.

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:

$$U_{ij} = -J_{ij} \vec{m}_i \cdot \vec{m}_j \quad (632)$$

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!*

The energetics of this are illustrated in figure 110. 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:

$$\vec{m}_e = \frac{e}{2m_e} \vec{s} \quad (633)$$

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

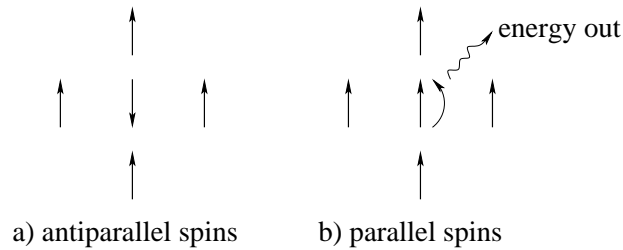


Figure 110: A cluster of five magnetic moments (spins) is illustrated with the central spins in two possible configurations. When the central spin is antiparallel to the four surrounding spins, it has potential energy $U_a = +4Jm^2$ in a suitable system of units. When it lines up parallel to the four surrounding spins, it's energy is $U_p = -4Jm^2$.

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*⁸¹.

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 potential energy of interaction between the spins. In very rough terms, as soon as the energy $k_B T$ (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) 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 ??) to illustrate the *permanent* macroscopic magnetization of a bar magnet in the *absence* of an external field.

The Curie Temperature and Neel Temperature

The critical temperature for the paramagnetic-ferromagnetic transition is called the *Curie Temperature* after Madame Curie, who discovered it. The critical temperature for the related antiferromagnetic transition is called the *Neel Temperature* for similar reasons (no, not because Curie discovered it, think harder).

Physicists find the classic ferromagnetic phase transition to be very interesting because it is an

⁸¹Wikipedia: 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.

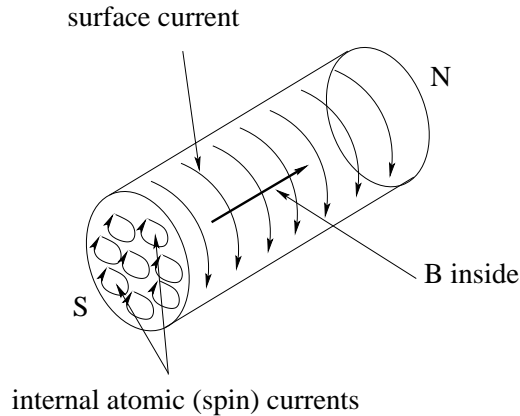


Figure 111: In a ferromagnet, the magnetic dipoles spontaneously align when cooled below a critical temperature. The resulting surface current transforms them into small “solenoids” with a non-dissipative surface current surrounding their interior volume and trapping magnetic flux that emerges from their north pole and flows to their south pole.

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 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 one⁸².

⁸²Sorry, 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, 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 boy 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...

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*⁸³ 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 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

⁸³Heh, heh. This is a pun, actually. If you don't get it now, you will. You will.

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 inexcusable, 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.

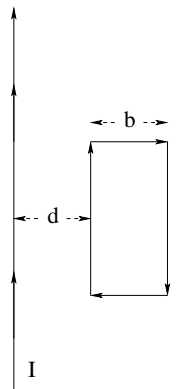
Homework for Week 8

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.

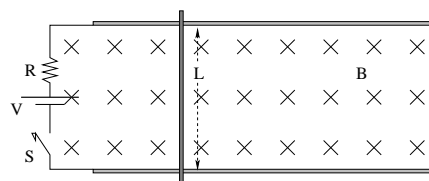


A long straight wire carries a current $I(t) = I_0 \sin(\omega t)$. A rectangular loop of wire with resistance R and dimensions $a \times b$ is a distance d away as shown. Find:

- the flux through the loop due to the wire;
- the mutual inductance M of the wire and loop;
- the induced voltage in the loop;
- the induced current in the loop;
- the force between the loop and the wire

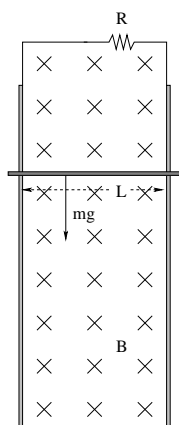
all as functions of time (where appropriate).

Problem 3.



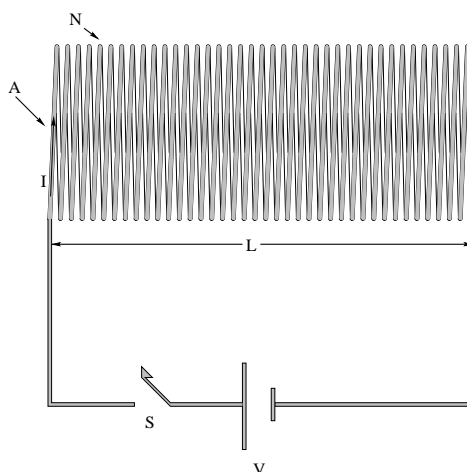
A rod of length L and mass m sits at rest on two frictionless conducting rails that sit in a plane perpendicular to a magnetic field as shown. At time $t = 0$ a switch S is closed connecting a voltage V that goes through a resistance R and the rod. The rod begins to move from a $x = 0$. Find:

- a) The terminal velocity of the rod (after the switch has been closed a very long time).
- b) The velocity of the rod as a function of time.
- c) The current in the loop as a function of time.

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. Find:

- a) The terminal velocity of the rod (after the rod has been falling for a very long time).
- b) The velocity of the rod as a function of time.
- c) The current in the loop as a function of time.

Problem 5.

- a) Find the self-inductance of the solenoid above that has N turns, length L , and circular radius A .

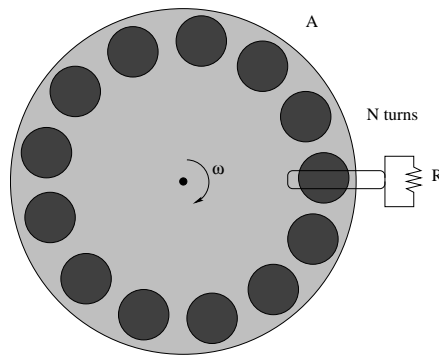
- b) Assuming that the conducting wire it is made of has radius a and resistivity ρ , find its resistance R .
- c) Find the current $I(t)$ in the circuit assuming that the switch S is closed at time $t = 0$.

Problem 6.

Complete the toroidal solenoid example begun for you above (see figure 103). Find the self-inductance L of a toroidal solenoid of N turns that has inner radius a , outer radius b , and height h .

Problem 7.

Complete the coaxial-cable example begun for you above (see figure 104). Find the high-frequency self-inductance *per unit length* of a coaxial cable with inner conductor radius a , outer conductor radius b .

Problem 8.

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 cross in front of a 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 flux $\phi = BA$.
- Assume that the flux of each magnet ramps up *linearly* from zero to ϕ and back down to zero in the time required for the magnet to swing past a loop.
- From this, estimate the induced voltage and current during the ramp up and ramp down phases. Plot them as a function of time for several cycles, assuming constant ω .
- Compute (and plot) the (effectively average) *power* during the ramp up and ramp down phases.
- **Advanced!** You should have gotten a power of the general form:

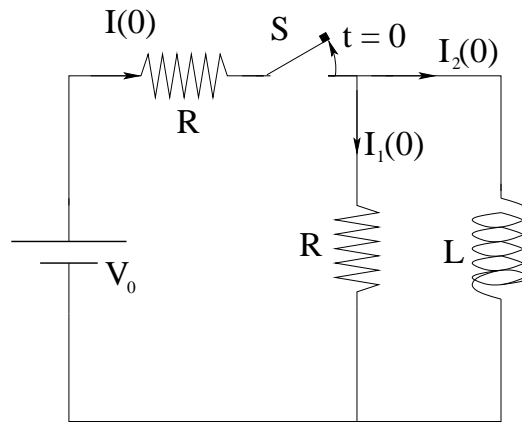
$$P = -C\omega^2 = \frac{dK}{dt}$$

for a constant C that depends on M , N , etc (not given as you are supposed to derive this).

This is the rate the kinetic energy of the car is being *reduced* to either heating the resistor or recharging the car's battery. As the kinetic energy is reduced, the car will *slow down*. Note well that $\omega = \frac{v}{r}$ where r is the radius of the tire. The kinetic energy of the car is $K = \frac{1}{2}m_c v^2 = \frac{1}{2}m_c \omega^2 r^2$.

So here's the challenge: Convert the expression for the power above into a differential equation for K , the kinetic energy of the car. Solve for the kinetic energy as a function of time, starting from an arbitrary initial value K_0 . Note well that the car would *never quite stop* if only magnetic braking were used (assuming that the model above is accurate even for very small speeds). Cars with magnetic brakes must *always* transition to friction brakes when the speed becomes small, because they exert less braking force and remove less energy as the car slows down!

In a car with magnetic brakes the loop would recharge a battery. In the next chapter we'll learn to treat oscillating voltages and power more accurately, but this estimate should suffice for the moment.

Problem 9.

In the circuit above, switch S has been closed for a very long time. At time $t = 0$ the switch is opened. Find:

- The currents $I(0)$, $I_1(0)$, and $I_2(0)$ at $t = 0$ at the instant before the switch is opened.
- Using Kirchhoff's voltage rule, find (derive) and solve the differential equation for $I_2(t)$. Draw a qualitative plot of this function.
- 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.

Week 9: Alternating Current Circuits

- **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:

$$\phi_m = \vec{B} \cdot N A \hat{n} = N B A \cos(\omega t) \quad (634)$$

$$V(t) = -\frac{d\phi_m}{dt} = N B A \omega \sin(\omega t) \quad (635)$$

We will from now on treat “arbitrary” harmonic alternating voltage sources as having the form:

$$V(t) = V_0 \sin(\omega t) \quad (636)$$

where of course we can introduce an arbitrary phase (corresponding to the choice of when we start our clock).

- 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): 209 is the potential difference between any two phases of a three-phase “Wye” main supply in the US where the pole voltages are 120 relative to ground:

$$\begin{aligned} V &= 120 \sin(\omega t) + 120 \sin(\omega t \pm 2\pi/3) \\ &= 240 \sin(\pi/3) \sin(\omega t \pm \pi/3) \\ &= 208 \sin(\omega t \pm \pi/3) \end{aligned} \quad (637)$$

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 internet⁸⁴.

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

⁸⁴Wikipedia: http://www.wikipedia.org/wiki/Mains_electricity. See also the many links in this article.

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 4: 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.

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!

- The reason for using such low frequencies is that AC does not flow uniformly through a conductor – it lies within an exponential distance of the *outer surface* of a conductor, a length called the *skin depth*. At 60 Hz this length is roughly 8.5 mm in copper; copper conductors “an inch in diameter” or more have relatively little current transmitted along their axis, where at 10 kHz (an arguably safer frequency) it is 0.66 mm in copper. Thicknesses comparable to the skin depth *increase the resistance* of a wire by effectively decreasing its cross-sectional area. 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.
- It is no exaggeration to state that this 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).
- There are two critical aspects of so-called alternating current (AC) that we will study in this course. The first is transformers and the electrical grid that delivers power to points distant from the generators with minimal loss. The second is the basis for signal processing electronics: the *LRC* band-pass circuit (or tank circuit) that can be used with rectifiers to build a simple amplitude-modulation (AM) radio. This circuit and its variants is ubiquitous in non-digital (and most digital) information processing devices.
- **The 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. 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:

$$V_s = N_s \frac{d\phi_m}{dt} \quad (638)$$

$$V_p = N_p \frac{d\phi_m}{dt} \quad (639)$$

or (taking the ratios of these two equations, in order)

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \quad (640)$$

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.

- Here's the trick of the power grid. 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 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 $I^2 R \approx 1$ 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 $I^2 R$ becomes your *friend* – delivering 10 kW at 100,000 V requires only 0.1 A and uses only 0.01 W heating the wire.

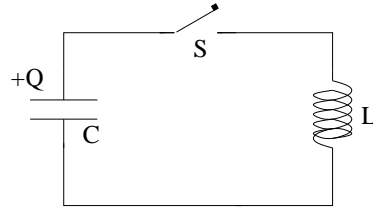
This is good for transmission, but bad for utilization. 100,000 volts can arc an appreciable distance through even *dry* air; that's why the insulators on high voltage transmission towers are so long! We'd hate to get electrocuted every time we changed a light bulb as power arced out of the socket through 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. Remember the crispy-fried squirrel story!

- 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). We use currents on the order of 1-20 Amps within the house, which is low enough that the resistive heating of the order of 30-50 meter long *household* supply lines remains low. Even “low” can waste a lot of heat! 12 gauge copper wire has a resistance of a bit less than 0.25 Ohms in 50 meters, wasting around 100 watts heating the wire all along its length when one draws 20 Amps of current (and reducing the line voltage available to the ~2000 watt appliance at the end that is drawing all of that power by roughly 5%). Personally, I prefer to do primary runs in household wiring with the even thick 10 gauge wire (and not to use the thinner 14 gauge wire *at all* to minimize heat loss in the household wiring. As you can see, though, you can easily waste anywhere from 1% to 5% of your energy bill simply heating the space inside your walls!
- **Non-driven LC circuit:** In the figure above, 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:

$$\frac{Q}{C} - L \frac{dI}{dt} = 0 \quad (641)$$

where

$$I = -\frac{dQ}{dt} \quad (642)$$

Figure 112: Undriven LC circuit

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:

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = 0 \quad (643)$$

We recognize this as the differential equation for a *harmonic oscillator*! To solve it, we “guess”⁸⁵:

$$Q(t) = Q_0 e^{\alpha t} \quad (644)$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{1}{LC} = 0 \quad (645)$$

We solve for:

$$\alpha = \pm i \sqrt{\frac{1}{LC}} = \pm i\omega_0 \quad (646)$$

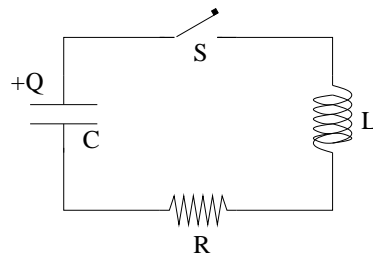
and get:

$$Q(t) = Q_{0+} e^{+i\omega_0 t} + Q_{0-} e^{-i\omega_0 t} \quad (647)$$

or (taking the real part and using the initial conditions):

$$Q(t) = Q_0 \cos(\omega_0 t) \quad (648)$$

- **Non-driven LRC circuit:** In the figure above, the capacitor C on the left is initially charged

Figure 113: Undriven LRC circuit

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:

$$\frac{Q}{C} - L \frac{dI}{dt} - IR = 0 \quad (649)$$

where

$$I = -\frac{dQ}{dt} \quad (650)$$

⁸⁵Not really.

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:

$$\frac{d^2 Q}{dt^2} + \frac{R}{L} \frac{dQ}{dt} + \frac{Q}{LC} = 0 \quad (651)$$

We recognize this as the differential equation for a *damped harmonic oscillator*. To solve it, we “guess”⁸⁶:

$$Q(t) = Q_0 e^{\alpha t} \quad (652)$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{R}{L} \alpha + \frac{1}{LC} = 0 \quad (653)$$

We solve for:

$$\begin{aligned} \alpha &= -\frac{R}{2L} \pm \frac{\sqrt{\left(\frac{R}{L}\right)^2 - \frac{4}{LC}}}{2} \\ &= -\frac{R}{2L} \pm i\omega_0 \sqrt{1 - \frac{R^2 C}{4L}} \\ &= -\frac{R}{2L} \pm i\omega_0 \sqrt{1 - \frac{\tau_C}{4\tau_L}} \\ &= -\frac{R}{2L} \pm i\omega' \end{aligned} \quad (654)$$

where $\tau_L = L/R$ $\tau_C = RC$, $\omega' = \omega_0 \sqrt{1 - \frac{\tau_C}{4\tau_L}}$, and our final solution looks like:

$$Q(t) = Q_0 e^{\frac{-Rt}{2L}} \cos(\omega' t) \quad (655)$$

(after we choose the real part of the complex exponential and use the initial conditions).

From this we can easily find the current through and voltage across all of the elements of the circuit. Finally, given the current and voltages it is easy to show that energy is conserved, that the initial energy stored in the capacitor exactly balances the energy consumed in the resistor as $t \rightarrow \infty$.

- **AC voltage across a resistance R :**

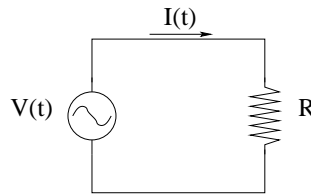


Figure 114: AC voltage across R

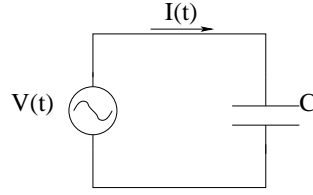
We use Kirchhoff's voltage rule and Ohm's Law to get:

$$V_0 \sin(\omega t) - IR = 0 \quad (656)$$

or

$$I_R(t) = \frac{V_0}{R} \sin(\omega t) \quad (657)$$

and we see that the current is *in phase* with the voltage drop across a resistor.

Figure 115: AC voltage across C

- **AC voltage across a capacitance C :**

We use Kirchhoff's voltage rule and the definition of capacitance to get:

$$V_0 \sin(\omega t) - \frac{Q}{C} = 0 \quad (658)$$

We can solve for $Q(t)$:

$$Q(t) = CV_0 \sin(\omega t) \quad (659)$$

Finally, we note that:

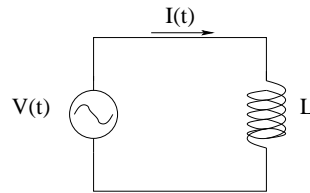
$$\begin{aligned} I_C(t) &= \frac{dQ(t)}{dt} = (\omega C)V_0 \cos(\omega t) \\ &= (\omega C)V_0 \sin(\omega t + \pi/2) = I_0 \sin(\omega t + \pi/2) \end{aligned} \quad (660)$$

where

$$I_0 = (\omega C)V_0 = \frac{V_0}{\chi_C} \quad (661)$$

We see that the current is $\pi/2$ *ahead in phase* of the voltage drop across the capacitor. We will actually usually use this the other way around and note that the voltage drop across the capacitor is $\pi/2$ *behind* the current through it. We call the quantity $\chi_C = \frac{1}{\omega C}$ (which clearly has the units of Ohms) the *capacitive reactance*, the “resistance” of a capacitor to alternating voltages.

- **AC voltage across an inductance L :**

Figure 116: AC voltage across L

We use Kirchhoff's voltage rule and the definition of capacitance to get:

$$V_0 \sin(\omega t) - L \frac{dI}{dt} = 0 \quad (662)$$

We can solve for $dI(t)$:

$$dI = \frac{V_0}{L} \sin(\omega t) dt \quad (663)$$

⁸⁶Not really.

We integrate both sides to get:

$$\begin{aligned} I_L(t) &= \int \frac{V_0}{L} \sin(\omega t) dt \\ &= \int \frac{V_0}{\omega L} \sin(\omega t) \omega dt \\ &= \frac{V_0}{\omega L} \cos(\omega t) \end{aligned} \quad (664)$$

$$= \frac{V_0}{\omega L} \sin(\omega t - \pi/2) \quad (665)$$

$$= I_0 \sin(\omega t - \pi/2) \quad (666)$$

$$(667)$$

where

$$I_0 = \frac{V_0}{\omega L} = \frac{V_0}{\chi_L} \quad (668)$$

We see that the current is $\pi/2$ *behind in phase* of the voltage drop across the inductor. We will actually usually use this the other way around and note that the voltage drop across the inductor is $\pi/2$ *ahead* of the current through it. We call the quantity $\chi_L = \omega L$ (which clearly has the units of Ohms) the *inductive reactance*, the “resistance” of an inductor to alternating voltages.

- **The Series LRC Circuit:** We apply Kirchhoff’s voltage/loop rule to this circuit and get:

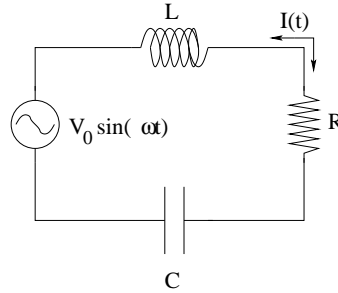


Figure 117: A series LRC (tank) circuit.

$$V_0 \sin(\omega t) - L \frac{dI}{dt} - RI - \frac{Q}{C} = 0 \quad (669)$$

or

$$V_L + V_R + V_C = V_0 \sin(\omega t) \quad (670)$$

or

$$\frac{d^2 Q}{dt^2} + \frac{R}{L} \frac{dQ}{dt} + \frac{1}{LC} Q = \frac{V_0}{L} \sin(\omega t) \quad (671)$$

There are a number of ways to solve this second order, linear, *inhomogeneous* ordinary differential equation. We will first show a simple one that relies on a “guess”, then we will show how if we use complex exponentials we really don’t have to guess.

Our goal will be to solve for all voltage drops, the current in the circuit, the power delivered to each circuit element and the entire circuit as a whole – pretty much everything.

The first thing to note is that if we find at least one “particular” solution $Q_p(t)$ to the inhomogeneous ODE, we can construct a new solution by adding *any* solution to the *homogeneous* ODE (the undriven LRC circuit solved above) and still get a solution. That is, a general solution can be written:

$$Q(t) = Q_p(t) + Q_h(t) \quad (672)$$

Note that the solution to the homogeneous ODE *decays in time* exponentially. It is a *transient* contribution to the overall solution and after many lifetimes $\tau_L = R/L$ it will generally be negligible.

The remaining particular part is therefore called the *steady state* part of the solution, and it persists indefinitely, as long as the driving voltage remains turned on. We expect that the time dependence of the steady state solution be *harmonic* (like the applied voltage) and to have the *same frequency* as the applied voltage. However, there is no particular reason to expect the charge Q to be *in phase* with the applied voltage.

We will find it slightly more convenient to work at first with the current I than the charge Q – we can always find $Q(t)$ (or V_C) by integration and V_L by differentiation – although when we go to a complex formulation it won't matter. If we make the *guess*:

$$I(t) = I_0 \sin(\omega t - \phi) \quad (673)$$

then solving the problem is easy⁸⁷. We begin by noting the voltage drops across all three circuit elements in terms of $I(t)$:

$$V_R = I_0 R \sin(\omega t - \phi) \quad (674)$$

$$V_L = I_0 \chi_L \sin(\omega t - \phi + \pi/2) \quad (675)$$

$$V_C = I_0 \chi_C \sin(\omega t - \phi - \pi/2) \quad (676)$$

or

$$\begin{aligned} I_0 R \sin(\omega t - \phi) &+ I_0 \chi_L \sin(\omega t - \phi + \pi/2) \\ &+ I_0 \chi_C \sin(\omega t - \phi - \pi/2) = V_0 \sin(\omega t) \end{aligned} \quad (677)$$

Our goal, then, is to find values of I_0 and ϕ for which this equation is *true*. This is quite simple. Suppose I use a *phasor diagram* to add the trig functions graphically: The y -components of

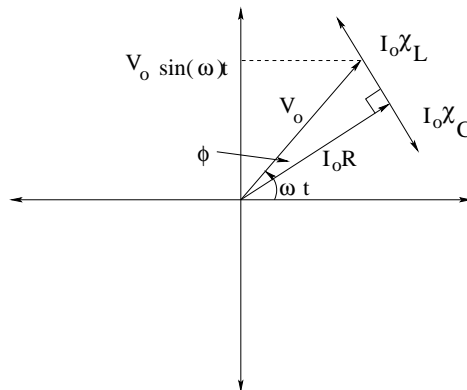
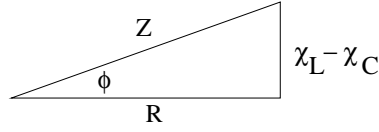


Figure 118: A phasor diagram for the LRC circuit.

the phasors on the diagram that are proportional to I_0 must add up to produce $V_0 \sin(\omega t)$, and this *must be true* if we add up the phasors as shown, taking advantage of our knowledge of the phase of the voltage drop across the various elements relative to the current through those elements.

If we let $V_0 = I_0 Z$ where Z is called the *impedance* of the circuit, we can *cancel* the I_0 and get the following triangle for the impedance: From this triangle we can easily see that:

⁸⁷This isn't really a guess. If we were to solve the differential equation "properly" using fourier transforms and using a complex exponential source $V_0 e^{i\omega t}$ we would discover that the complex solution for the current has a complex amplitude and phase determined from an algebraic equation. We are simply making the guess here because many students don't know enough math yet to handle this approach, although this may change in some future edition of this book.

Figure 119: The impedance diagram for the *LRC* circuit.

$$Z = \sqrt{R^2 + (\chi_L - \chi_C)^2} \quad (678)$$

so that

$$I_0 = \frac{V_0}{Z} \quad (679)$$

and

$$\phi = \tan^{-1} \left(\frac{\chi_L - \chi_C}{R} \right) \quad (680)$$

• **The Parallel LRC Circuit:**

The parallel LRC circuit is actually much *simpler* than the series as far as understanding the solution is concerned. This is because the *same* voltage drop $V_0 \sin(\omega t)$ occurs across all *three* components, and so we can just write down the currents through each component using the elementary single-component rules above:

$$I_R = \frac{V_0}{R} \sin(\omega t) \quad (681)$$

$$I_L = \frac{V_0}{\chi_L} \sin(\omega t - \pi/2) \quad (682)$$

$$I_C = \frac{V_0}{\chi_C} \sin(\omega t + \pi/2) \quad (683)$$

Note well that we use the rules we derived where the current through the inductor is $\pi/2$ *behind* the voltage (which is therefore $\pi/2$ *ahead* of the current) and vice versa for the capacitor. To find the total current provided by the voltage, we simply add these three currents according to Kirchhoff's junction rule. Of course, we are adding three trig functions with different relative phases, so we once again must accomplish this with suitable phasors:

$$\begin{aligned} I_{\text{tot}} &= \frac{V_0}{R} \sin(\omega t) + \frac{V_0}{\chi_L} \sin(\omega t - \pi/2) + \frac{V_0}{\chi_C} \sin(\omega t + \pi/2) \\ &= \frac{V_0}{Z} \sin(\omega t - \phi) \\ &= I_0 \sin(\omega t - \phi) \end{aligned} \quad (684)$$

In this expression, a bit of contemplation should convince you that the impedance Z for this circuit is given by the entirely reasonable:

$$\frac{1}{Z} = \sqrt{\left(\frac{1}{R^2} + \left(\frac{1}{\chi_C} - \frac{1}{\chi_L} \right)^2 \right)} \quad (685)$$

which we recognize as the phasor equivalent of the familiar rule for reciprocal addition of resistances in parallel, and:

$$\begin{aligned} \phi &= \tan^{-1} \left(\frac{\frac{1}{\chi_C} - \frac{1}{\chi_L}}{\frac{1}{R}} \right) \\ &= \tan^{-1} \left(\frac{RC(\omega^2 - \omega_0^2)}{\omega} \right) \end{aligned} \quad (686)$$

for the phase.

Resonance for this circuit is a bit unusual – it is the frequency $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$ as before, but now $\frac{1}{Z}$ is *largest* at resonance and the current increases away from resonance. The power delivered to the resistance no longer depends on L or C and only depends on the frequency as:

$$P_R = \frac{V_0^2 \sin^2(\omega t)}{R} \quad (687)$$

so that the average power delivered to the circuit is:

$$\langle P \rangle = \langle P_R \rangle = \frac{V_0^2}{2R} \quad (688)$$

independent of frequency altogether. Away from resonance, one simply generates a large (but irrelevant) current in either L (for low frequencies) or C (for high frequencies) that is out of phase with the voltage and hence dissipates zero average power per cycle.

9.1: 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:

$$\phi_m = \vec{B} \cdot N A \hat{n} = N B A \cos(\omega t) \quad (689)$$

$$V(t) = -\frac{d\phi_m}{dt} = N B A \omega \sin(\omega t) \quad (690)$$

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:

$$V(t) = V_0 \sin(\omega t) \quad (691)$$

where of course we can introduce an arbitrary phase (corresponding to the choice of when we start our clock). In this expression, remember that:

$$\omega = 2\pi f = \frac{2\pi}{T} \quad (692)$$

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:

$$V(t) = A(t) \sin(\omega t) \quad (693)$$

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:

$$V(t) = V_0 e^{i\omega t} \quad (694)$$

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

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 5: 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.

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 alternating 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!*

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): 209 is the potential difference between any two phases of a three-phase “Wye” main supply in the US where the pole voltages are 120 relative to ground:

$$\begin{aligned}
 V &= 120 \sin(\omega t) + 120 \sin(\omega t \pm 2\pi/3) \\
 &= 240 \sin(\pi/3) \sin(\omega t \pm \pi/3) \\
 &= 208 \sin(\omega t \pm \pi/3)
 \end{aligned} \tag{695}$$

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 internet⁸⁸.

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

⁸⁸Wikipedia: http://www.wikipedia.org/wiki/Mains_electricity. See also the many links in this article.

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 proceeds 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

$$R = \frac{\rho L}{A} \quad (696)$$

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.

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.

The 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. 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

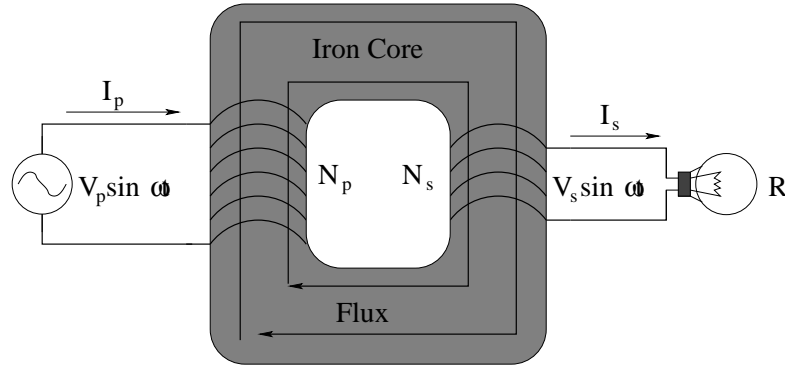


Figure 120: A transformer transforms voltage V_1 into a new voltage V_2 , for time-varying (usually sinusoidal) voltages only.

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:

$$V_s = N_s \frac{d\phi_m}{dt} \quad (697)$$

$$V_p = N_p \frac{d\phi_m}{dt} \quad (698)$$

or (taking the ratios of these two equations, in order)

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \quad (699)$$

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):

$$P_p = V_p I_p = V_s I_s = P_s \quad (700)$$

or, if we use the fact that $V_s = V_p N_s / N_p$ and divide a couple of times, we find that:

$$I_s = \frac{N_p}{N_s} I_p \quad (701)$$

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 if 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.

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!

At the same time, a young man named Nikola Tesla⁸⁹, who *was* a competent mathematician and who had worked for Edison for a while before he resigned claiming (correctly) that he was undervalued by Edison, who cheated him out of a promised payment of \$50,000. He realized that the secret to the economical transmission of power was the use of high voltages (and correspondingly low currents) in the transmission process, something that is only possible if one uses alternating current (AC) and transformers like the one schematized above. Tesla quit working for Edison (and General Electric) and ultimately went to work for Westinghouse, that gradually prospered on the basis of his new scheme. This was the so-called *War of the Currents*⁹⁰.

Edison lost (although Tesla never really benefitted much from his victory, being cursed with bad luck that seemed to guarantee that he would never become rich from his cornucopia of enormously valuable inventions. Tesla was dead on correct – Edison's solution was no solution at all and could never have supported the centralized generation and distribution of electrical power that is the fundamental basis of modern civilization with its vast and distributed productivity and its unprecedented degree of personal comfort and information access, all enabled by mass-produced electrical power distributed by Tesla's solution. We absolutely need to learn, and understand, this solution as it is of paramount importance *today*, some 130 years later, as we struggle to convert

⁸⁹Wikipedia: 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 worked for Edison redesigning Edison's DC generators (and was cheated of the promised payment), invented the Tesla coil, polyphase generators and motors, 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 strongly recommend reading at least the wikipedia article on him.

⁹⁰Wikipedia: http://www.wikipedia.org/wiki/War_of_Currents.

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 of miles* from those locations to where it will be consumed.

So here's the trick of the power grid, Tesla'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 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 $I^2 R \approx 1$ 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 $I^2 R$ 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 recently 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 flick a switch on a humid day, you risked being electrocuted by a 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). We use currents on the order of 1-20 Amps within the house, which is low enough that the resistive heating of the order of 30-50 meter long *household* supply lines remains low. Even "low" can waste a lot of heat! 12 gauge copper wire has a resistance of a bit less than 0.25 Ohms in 50 meters, wasting around 100 watts heating the wire all along its length when one draws 20 Amps of current (and reducing the line voltage available to the ~2000 watt appliance at the end that is drawing all of that power by roughly 5%). Personally, I prefer to do primary runs in household wiring with the even thicker 10 gauge wire (and not to use the thinner 14 gauge wire *at all* to minimize heat loss in the household wiring. 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!

All of this will make sense when you work out the algebra for yourself. One of the homework

problems has you do this very thing. 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 associate 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.

9.2: 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 piston 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 repetition 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.

Non-driven LC circuit

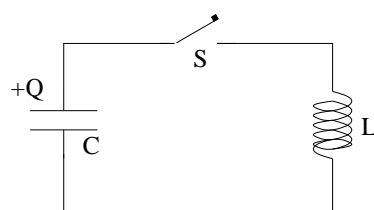


Figure 121: Undriven LC circuit

In the figure above, 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:

$$\frac{Q}{C} - L \frac{dI}{dt} = 0 \quad (702)$$

where

$$I = -\frac{dQ}{dt} \quad (703)$$

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:

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = 0 \quad (704)$$

We recognize this as the differential equation for a *harmonic oscillator*! To solve it, we “guess”⁹¹:

$$Q(t) = Q_0 e^{\alpha t} \quad (705)$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{1}{LC} = 0 \quad (706)$$

We solve for:

$$\alpha = \pm i \sqrt{\frac{1}{LC}} = \pm i\omega_0 \quad (707)$$

and get:

$$Q(t) = Q_{0+} e^{+i\omega_0 t} + Q_{0-} e^{-i\omega_0 t} \quad (708)$$

or (taking the real part and using the initial conditions):

$$Q(t) = Q_0 \cos(\omega_0 t) \quad (709)$$

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:

$$U_C(t) = \frac{Q(t)^2}{2C} \quad (710)$$

This energy *over time* oscillates between the capacitor and the energy in the inductor:

$$U_L(t) = \frac{1}{2} L I(t)^2 \quad (711)$$

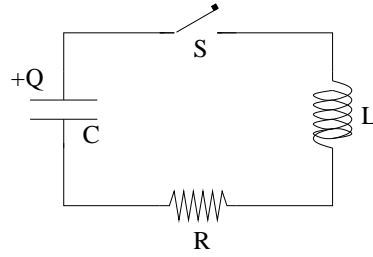
Show that the sum of these two energies is a constant, and that the constant equals the initial energy in the capacitor! This is precisely analogous to what happens to the conserved total energy as it oscillates between potential energy in a spring and kinetic energy of motion of the mass in a harmonic oscillator.

Non-driven LRC circuit

In the figure above, 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:

$$\frac{Q}{C} - L \frac{dI}{dt} - IR = 0 \quad (712)$$

⁹¹Not really.

Figure 122: Undriven LRC circuit

where

$$I = -\frac{dQ}{dt} \quad (713)$$

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:

$$\frac{d^2 Q}{dt^2} + \frac{R}{L} \frac{dQ}{dt} + \frac{Q}{LC} = 0 \quad (714)$$

We recognize this as the differential equation for a *damped harmonic oscillator*. To solve it, we “guess”⁹²:

$$Q(t) = Q_0 e^{\alpha t} \quad (715)$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{R}{L}\alpha + \frac{1}{LC} = 0 \quad (716)$$

We solve for:

$$\begin{aligned} \alpha &= -\frac{R}{2L} \pm \frac{\sqrt{\left(\frac{R}{L}\right)^2 - \frac{4}{LC}}}{2} \\ &= -\frac{R}{2L} \pm i\omega_0 \sqrt{1 - \frac{R^2 C}{4L}} \\ &= -\frac{R}{2L} \pm i\omega_0 \sqrt{1 - \frac{\tau_L}{4\tau_R}} \\ &= -\frac{R}{2L} \pm i\omega' \end{aligned} \quad (717)$$

where $\tau_L = R/L$, $\tau_C = 1/RC$, $\omega' = \omega_0 \sqrt{1 - \frac{\tau_L}{4\tau_R}}$, and our final solution looks like:

$$Q(t) = Q_0 e^{-\frac{Rt}{2L}} \cos(\omega' t) \quad (718)$$

(after we choose the real part of the complex exponential and use the initial conditions).

Note well the analogy with the damped, undriven harmonic oscillator.

From this we can easily find the current through and voltage across all of the elements of the circuit. Finally, given the current and voltages it is easy to show that energy is conserved, that the initial energy stored in the capacitor exactly balances the energy consumed in the resistor as $t \rightarrow \infty$. This is again left as an exercise – the more of this that you work out on your own (it is quite easy – compute the *power* delivered to each circuit element over time and integrate over time to find the total energy consumed by the resistor or residual in the capacitor or inductor) the better you will learn it.

⁹²Not really.

To go on, we need to introduce a classical harmonic oscillating voltage like that produced by an AC generator. 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.

A Harmonic AC Voltage Across a Resistance R

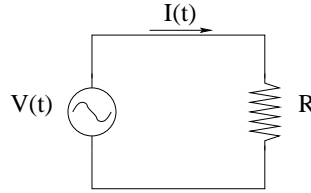


Figure 123: AC voltage across R

Consider the circuit diagram in figure 123, portraying an alternating voltage $V(t) = V_0 \sin(\omega t)$ placed across a resistance R . Applying Kirchhoff's voltage/loop rule and Ohm's Law to the circuit loop, we get the following equation of motion for the circuit:

$$V_0 \sin(\omega t) - IR = 0 \quad (719)$$

or (solving for the desired current):

$$I_R(t) = \frac{V_0}{R} \sin(\omega t) \quad (720)$$

and we see that the current is *in phase* with the voltage drop across a resistor.

A Harmonic AC Voltage Across a Capacitance C

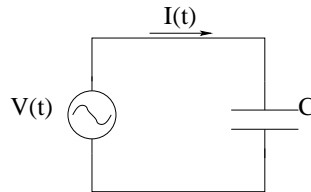


Figure 124: AC voltage across CR

Proceeding the exact same way, we use Kirchhoff's voltage rule and the definition of capacitance to get an equation of motion:

$$V_0 \sin(\omega t) - \frac{Q}{C} = 0 \quad (721)$$

We wish to find $I_C(t)$, the current through the capacitor. To get it, first we solve for $Q(t)$:

$$Q(t) = CV_0 \sin(\omega t) \quad (722)$$

and then we differentiate (and use the trigonometric identity $\cos(\theta) = \sin(\theta + \pi/2)$ to express the result in terms of the original harmonic function and a phase):

$$\begin{aligned} I_C(t) &= \frac{dQ(t)}{dt} = (\omega C)V_0 \cos(\omega t) \\ &= (\omega C)V_0 \sin(\omega t + \pi/2) \end{aligned} \quad (723)$$

We observe in this result a quantity that behaves like the “resistance” of the capacitor in an AC circuit, regulating the magnitude of the current as the frequency changes much the way a resistor would as it increases or decreases. We will give this quantity its own name – the *capacitive reactance* of the capacitor at the angular frequency ω – and define it to be:

$$\chi_C = \frac{1}{\omega C} \quad (724)$$

Note well that the units of χ_C are *ohms*.

Using the capacitive reactance, the peak current in the circuit takes on a more familiar form:

$$I_0 = (\omega C)V_0 = \frac{V_0}{\chi_C} \quad (725)$$

so that

$$I_C(t) = I_0 \sin(\omega t + \pi/2) \quad (726)$$

We see that the current is $\pi/2$ *ahead in phase* of the voltage drop across the capacitor. We will actually usually use this in series AC circuits with capacitors the other way around and note that the voltage drop across the capacitor is $\pi/2$ *behind* the current through it.

A Harmonic AC Voltage Across an Inductance L

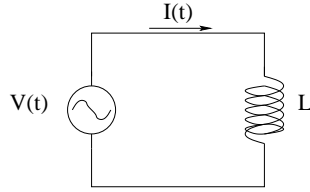


Figure 125: AC voltage across L

We repeat this process one more time for an inductance L . The methodology is basically the same: We use Kirchhoff's voltage rule and the definition of capacitance to get:

$$V_0 \sin(\omega t) - L \frac{dI}{dt} = 0 \quad (727)$$

This time we solve for $dI(t)$:

$$dI = \frac{V_0}{L} \sin(\omega t) dt \quad (728)$$

and integrate both sides to get:

$$\begin{aligned} I_L(t) &= \int \frac{V_0}{L} \sin(\omega t) dt \\ &= \int \frac{V_0}{\omega L} \sin(\omega t) \omega dt \\ &= \frac{V_0}{\omega L} \cos(\omega t) \end{aligned} \quad (729)$$

$$= \frac{V_0}{\omega L} \sin(\omega t - \pi/2) \quad (730)$$

We define the *inductive reactance*:

$$\chi_L = \omega L \quad (731)$$

(in ohms once again) so that:

$$I_L(t) = I_0 \sin(\omega t - \pi/2) \quad (732)$$

with a peak current given by:

$$I_0 = \frac{V_0}{\omega L} = \frac{V_0}{\chi_L} \quad (733)$$

We see that the current is $\pi/2$ *behind in phase* of the voltage drop across the inductor. As before, in series circuits we will actually use this the other way around, considering the voltage drop across and inductor in a series circuit as being $\pi/2$ *ahead* of the (otherwise specified) current through it. Let's see how this works.

The Series LRC Circuit

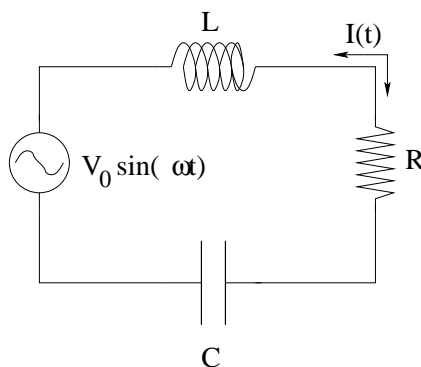


Figure 126: A *LRC* (tank) circuit.

In figure 126 above we see a *series LRC* circuit. To analyze it, we apply Kirchhoff's voltage/loop rule to the circuit and get the equation of motion:

$$V_0 \sin(\omega t) - L \frac{dI}{dt} - RI - \frac{Q}{C} = 0 \quad (734)$$

or

$$V_L + V_R + V_C = V_0 \sin(\omega t) \quad (735)$$

or

$$\frac{d^2 Q}{dt^2} + \frac{R}{L} \frac{dQ}{dt} + \frac{1}{LC} Q = \frac{V_0}{L} \sin(\omega t) \quad (736)$$

There are a number of ways to solve this second order, linear, *inhomogeneous* ordinary differential equation. We will first show a simple one that relies on a “guess”, then we will show how if we use complex exponentials we really don't have to guess. In fact, if we use complex exponentials for *everything* associated with electrical circuits and harmonic oscillation we don't really need to guess, but we do need to know more math than a student of introductory physics might initially know. Wise students will view this as an open invitation to *learn more math* to make the physics easier.

Our goal will be to solve for all voltage drops, the current in the circuit, the power delivered to each circuit element and the entire circuit as a whole – pretty much everything.

The first thing to note is that if we find at least one “particular” solution $Q_p(t)$ to the inhomogeneous ODE, we can construct a new solution by adding *any* solution to the *homogeneous* ODE (the undriven *LRC* circuit solved above) and still get a solution. That is, a general solution can be written:

$$Q(t) = Q_p(t) + Q_h(t) \quad (737)$$

Note that the solution $Q_h(t)$ to the homogeneous ODE (equation 718) *decays in time* exponentially. It is a *transient* contribution to the overall solution and after many lifetimes $\tau_L = R/L$ it will generally be negligible.

The remaining particular solution $Q_p(t)$ is therefore called the *steady state* part of the solution, and it persists indefinitely, as long as the driving voltage remains turned on. We expect that the time dependence of the steady state solution be *harmonic* (like the applied voltage) and to have the *same frequency* as the applied voltage. However, there is no particular reason to expect the charge Q to be *in phase* with the applied voltage.

We will find it slightly more convenient to work at first with the current I than the charge Q – we can always find $Q(t)$ (or V_C) by integration and V_L by differentiation – although when we go to a complex formulation it won't matter. If we make the *guess*:

$$I(t) = I_0 \sin(\omega t - \phi) \quad (738)$$

then solving the problem is easy⁹³. We begin by noting the voltage drops across all three circuit elements in terms of $I(t)$ (where we use the rules we derived above *backwards* as we are given the current and seek the voltage):

$$V_R = I_0 R \sin(\omega t - \phi) \quad (739)$$

$$V_L = I_0 \chi_L \sin(\omega t - \phi + \pi/2) \quad (740)$$

$$V_C = I_0 \chi_C \sin(\omega t - \phi - \pi/2) \quad (741)$$

or (substituting into Kirchoff's loop rule for the voltage):

$$\begin{aligned} I_0 R \sin(\omega t - \phi) &+ I_0 \chi_L \sin(\omega t - \phi + \pi/2) \\ &+ I_0 \chi_C \sin(\omega t - \phi - \pi/2) = V_0 \sin(\omega t) \end{aligned} \quad (742)$$

Our goal, then, is to find values of I_0 and ϕ for which this equation is *true*. This is quite simple. Suppose I use a *phasor diagram* to add the trig functions graphically: The y -components of the

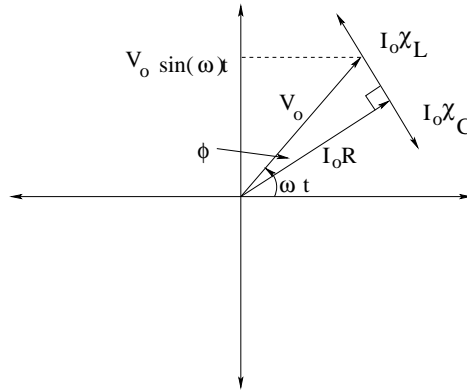


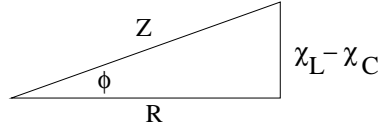
Figure 127: A phasor diagram for the series *LRC* circuit.

phasors on the diagram that are proportional to I_0 must add up to produce $V_0 \sin(\omega t)$, and this *must be true* if we add up the phasors as shown, taking advantage of our knowledge of the phase of the voltage drop across the various elements relative to the current through those elements.

If we let $V_0 = I_0 Z$ where Z is called the *impedance* of the circuit, we can *cancel* the I_0 and get the following triangle for the impedance: From this triangle we can easily see that:

$$Z = \sqrt{R^2 + (\chi_L - \chi_C)^2} \quad (743)$$

⁹³This isn't really a guess. If we were to solve the differential equation "properly" using fourier transforms and using a complex exponential source $V_0 e^{i\omega t}$ we would discover that the complex solution for the current has a complex amplitude and phase determined from an algebraic equation. We are simply making the guess here because many students don't know enough math yet to handle this approach, although this may change in some future edition of this book.

Figure 128: The impedance diagram for the *LRC* circuit.

so that

$$I_0 = \frac{V_0}{Z} \quad (744)$$

and

$$\phi = \tan^{-1} \left(\frac{\chi_L - \chi_C}{R} \right) \quad (745)$$

Power in a Series *LRC* Circuit

Power in this circuit is worth a section all its own, as understanding power delivery to the circuit is essential to the understanding of *why* this circuit is useful. The series *LRC* circuit functions as a *band pass filter* to an applied harmonic voltage. Basically, it only allows a large current to flow (and deliver power to the circuit) when the frequency of the applied voltage is (nearly) the *same* as the *resonant frequency for the circuit*:

$$\omega_0 = \sqrt{\frac{1}{LC}} \quad (746)$$

obtained above for the undriven *LC* circuit. For frequencies far from resonance, the current delivered and power dissipated in the circuit rapidly goes to zero. Let us understand this.

First we consider the power delivered to or by each circuit element. The power delivered *by* the voltage *to* the circuit is just:

$$P(t) = V(t)I(t) = V_0 \sin(\omega t) I_0 \sin(\omega t - \phi) \quad (747)$$

If we use the trig identity:

$$\sin(A - B) = \sin(A) \cos(B) - \cos(A) \sin(B) \quad (748)$$

(which can be trivially proven with complex exponentials) and $I_0 = V_0/Z$ we get:

$$P(t) = \frac{V_0^2}{Z} (\sin^2(\omega t) \cos(\phi) - \sin(\omega t) \cos(\omega t) \sin(\phi)) \quad (749)$$

We don't usually care about the *instantaneous* power delivered to the circuit (although there are very definitely exceptions, such as when the peak power is *much larger* than the average power, which can stress e.g. transformers that might be providing the power). If we time average this to obtain the *average* power we get:

$$\langle P(t) \rangle = P_{\text{av}} = \frac{V_0^2}{2Z} \cos(\phi) = \frac{V_0^2 R}{2Z^2} = \frac{1}{2} I_0^2 R \quad (750)$$

where we used the fact that the time average of the square of any harmonic function of time is $1/2$, the fact that $\cos(\phi) = R/Z$ from the impedance triangle above, and the fact that $I_0 = V_0/Z$. Note as well that the time average of $\sin(\omega t) \cos(\omega t)$ is *zero* (why?) so that the second term does not contribute.

Now consider each of the *other* circuit elements separately:

$$P_R(t) = V_R(t)I(t) = I_0^2 R \sin^2(\omega t - \phi) \quad (751)$$

$$P_L(t) = V_L(t)I(t) = I_0^2 \chi_L \sin(\omega t - \phi + \pi/2) \sin(\omega t - \phi) \quad (752)$$

$$P_C(t) = V_C(t)I(t) = I_0^2 \chi_C \sin(\omega t - \phi - \pi/2) \sin(\omega t - \phi) \quad (753)$$

Again we don't much care about the peak values, but the averages are important. The time averages of $\sin^2(\omega t - \phi)$ is $1/2$. The time average of $\sin(\omega t - \phi \pm \pi/2) \sin(\omega t - \phi)$ is zero (why?)! Thus the average power delivered to the circuit is delivered to the *resistor only*.

$$\langle P_R(t) \rangle = \frac{1}{2} I_0^2 R = \langle P(t) \rangle = P_{av} \quad (754)$$

from above. We see that the energy flowing in and out of the capacitor and inductor may be very large (if e.g. $I_0^2 \chi_L$ is large), but they *use* no energy and hence in every cycle the net energy they absorb from the circuit equals the net energy they return to the circuit!

Now let us consider P_{av} , understanding that it is delivered to the *resistor* (or circuit element such as an amplifier that behaves like a resistance in the circuit) and not the inductance or capacitance.

$$P_{av} = \frac{V_0^2 R}{2Z^2} = \frac{V_{rms}^2 R}{R^2 + (\chi_L - \chi_C)^2} = \frac{V_{rms}^2 R}{R^2 + (\omega L - \frac{1}{\omega C})^2} \quad (755)$$

where we have introduced the *root-mean square voltage*

$$V_{rms} = \frac{1}{\sqrt{2}} V_0 \quad (756)$$

as a form of the voltage that lets us drop the pesky factor of $1/2$ that frequently arises from averages in harmonic circuits (and leaves us with quantities that look more like their direct current counterparts, easier to remember).

We will often want to express this quantity in terms of *impedance* (which determines the current) and a convenient quantity (that we saw arose quite naturally above):

$$P_{av} = \frac{V_0^2 R}{Z^2} = \frac{1}{2} V_0 I_0 \cos \phi = V_{rms} I_{rms} P_f \quad (757)$$

where

$$P_f = \cos(\phi) = \frac{R}{Z} \quad (758)$$

is called the *power factor* of the circuit. When $P_f = 1$, $Z = R$ and the load is said to be *entirely resistive*. A lightbulb plugged into a wall is an example of a purely resistive load. When the power factor is different from one, in general the *peak* power delivered to the circuit is much greater than the *average* power, which means that the power supply has to deliver much larger peak voltages than you expect from the power rating of the appliance being used. This can in turn blow fuses or circuit breakers for a load that a circuit "should" be able to manage.

Let us consider the variation of the average power with frequency for

fixed circuit elements. The first thing to note is that power is obviously at a max when $\chi_L = \chi_C$:

$$\begin{aligned} \omega L &= \frac{1}{\omega C} \\ \omega^2 &= \frac{1}{LC} = \omega_0^2 \end{aligned} \quad (759)$$

where ω_0 is the *resonant frequency* of the circuit. At resonance the power delivered to the circuit is:

$$P_{av,max} = \frac{V_{rms}^2 R}{R^2} = \frac{V_{rms}^2}{R} \quad (760)$$

just as we would expect for a DC circuit. If we use peak instead of rms voltage, of course, we have to put back the factor of $1/2$:

$$P_{av,max} = \frac{V_0^2}{2R} \quad (761)$$

Next, let's factor the power into a slightly more convenient form:

$$\begin{aligned}
 P_{\text{av}} &= \frac{V_{\text{rms}}^2 R}{R^2 + (\omega L - \frac{1}{\omega C})^2} \\
 &= \frac{V_{\text{rms}}^2 R}{R^2 + \frac{L^2}{\omega^2} (\omega^2 - \frac{1}{LC})^2} \\
 &= \frac{V_{\text{rms}}^2 R \omega^2}{R^2 \omega^2 + L^2 (\omega^2 - \omega_0^2)^2}
 \end{aligned} \tag{762}$$

This function is plotted, for $L = C = V_0 = 1.0$ and several values of R (and hence Q) in figure 129. When $\omega \rightarrow 0$, $P_{\text{av}} \rightarrow 0$ like ω^2 . When $\omega \rightarrow \infty$, $P_{\text{av}} \rightarrow 0$ like $1/\omega^2$. In between, it clearly peaks at $\omega = \omega_0$, resonance, with a peak power as given above. We are thus almost ready to draw a *generic shape* for the *resonance curve*, the power delivered to the circuit as a function of frequency.

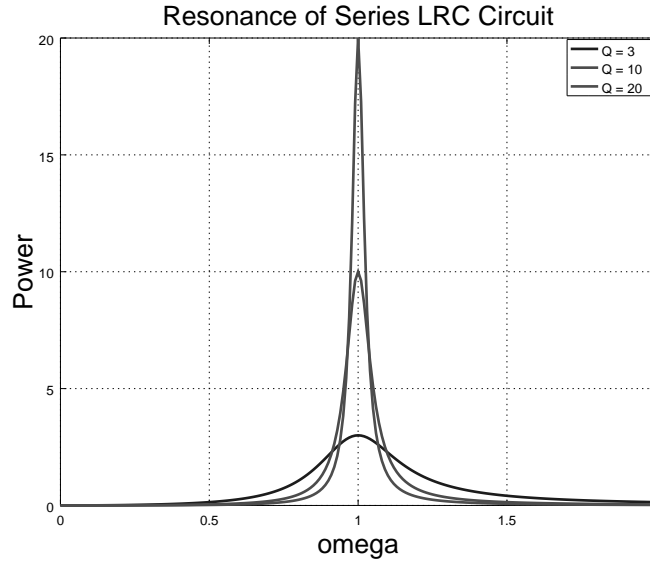


Figure 129: A typical series of resonance curves for $Q = 3, 10, 20$, plotted on a scale such that $\omega_0 = 1$: $L = C = 1.0$, and $R = 0.3333, 0.1, 0.05$.

To do so, however, we need to learn one last concept that is extremely useful in understanding the behavior of electrical band pass circuits: the *Q-factor* (quality factor) of the circuit. The *Q-factor* of a circuit is defined to be:

$$Q = \frac{\omega_0}{\Delta\omega} \tag{763}$$

where $\Delta\omega$ is the *full width* of the resonance curve at *half-maximum*. The *Q-factor* is a measure of the *sharpness* of the resonance. A circuit with a low *Q-factor* delivers significant power to the circuit for frequencies far from resonance (although asymptotically the power still vanishes at zero and infinity as given above). A circuit with a high *Q-factor* has a *sharply peaked resonance curve* that goes to zero *quickly* when ω is far from resonance – power is delivered to a circuit only for frequencies *very close to the resonance frequency*.

In the homework you will be asked to derive the relation:

$$Q = \frac{\omega_0 L}{R} = \frac{L}{\sqrt{LC}R} = \frac{1}{R} \sqrt{\frac{L}{C}} = \frac{\omega_0}{\Delta\omega} \tag{764}$$

In figure 129, you can see how decreasing R (with V_0 , L , and C all fixed at one) causes the resonance to *sharpen up* – become much narrower at half-max – at the same time it increases the maximum power delivered at peak dramatically.

In a later section we'll see at least one or two places one can *use* an series *LRC* circuit to do useful things, but first we have to study an even more useful circuit, the *parallel LRC* circuit.

The Parallel LRC Circuit

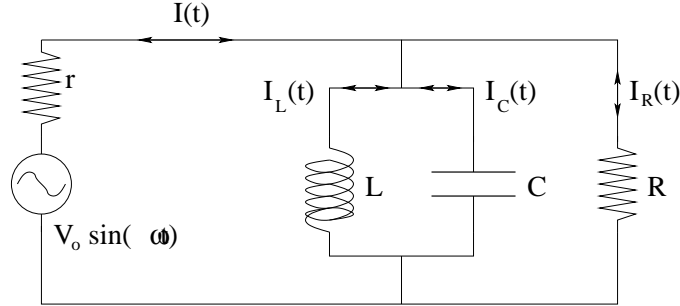


Figure 130: A parallel *LRC* circuit, with a voltage that has an “internal resistance” that limits its ability to deliver current. This circuit is ideal for the construction of a simple AM crystal radio.

The parallel *LRC* circuit drawn in figure 130 above is actually much *simpler* than the series as far as understanding the solution is concerned. In this figure we have added the internal resistance r of the power supply or antenna, as in the latter case especially the fact that the voltage cannot supply an infinite amount of power is essential to understanding how this circuit can be used to build a crystal radio. Note that we didn't bother doing this in the case of the series *LRC* circuit because the resistance R in that case was the *total* resistance from all sources in the single circuit loop.

It is simple to analyze because the *same* voltage drop $V_0 \sin(\omega t)$ occurs across all *three* components, and so we can just write down the currents through each component using the elementary single-component rules above:

$$I_R = \frac{V_0}{R} \sin(\omega t) \quad (765)$$

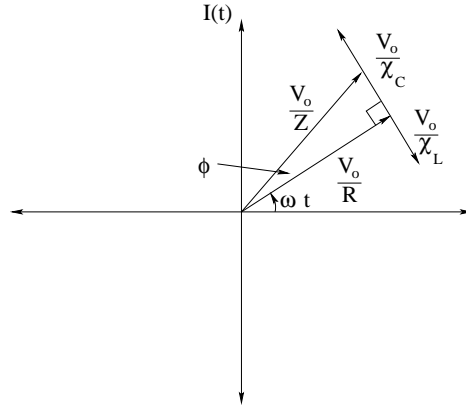
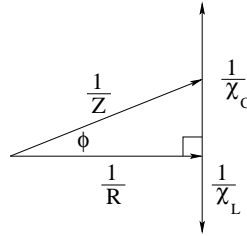
$$I_L = \frac{V_0}{\chi_L} \sin(\omega t - \pi/2) \quad (766)$$

$$I_C = \frac{V_0}{\chi_C} \sin(\omega t + \pi/2) \quad (767)$$

Note well that we use the rules we derived where the current through the inductor is $\pi/2$ *behind* the voltage (which is therefore $\pi/2$ *ahead* of the current) and vice versa for the capacitor. To find the total current provided by the voltage, we simply add these three currents according to Kirchhoff's junction rule. Of course, we are adding three trig functions with different relative phases, so we once again must accomplish this with suitable phasors:

$$\begin{aligned} I_{\text{tot}} &= \frac{V_0}{R} \sin(\omega t) + \frac{V_0}{\chi_L} \sin(\omega t - \pi/2) + \frac{V_0}{\chi_C} \sin(\omega t + \pi/2) \\ &= \frac{V_0}{Z} \sin(\omega t + \phi) \\ &= I_0 \sin(\omega t + \phi) \end{aligned} \quad (768)$$

As before, we can factor out the common V_0 and look at the resulting triangle addition of the *inverse* resistance and reactances to obtain a sum rule for the *inverse* impedance Z : From figure 132

Figure 131: A phasor diagram for the parallel LRC circuit.Figure 132: The impedance diagram for the parallel LRC circuit.

the pythagorean theorem immediately yields an expression for the inverse of the impedance:

$$\frac{1}{Z} = \sqrt{\left(\frac{1}{R^2} + \left(\frac{1}{X_C} - \frac{1}{X_L}\right)^2\right)} \quad (769)$$

which we recognize as the phasor equivalent of the familiar rule for reciprocal addition of resistances in parallel.

We similarly can easily evaluate the phase ϕ :

$$\begin{aligned} \phi &= \tan^{-1} \left(\frac{\frac{1}{X_C} - \frac{1}{X_L}}{\frac{1}{R}} \right) \\ &= \tan^{-1} \left(\frac{RC(\omega^2 - \omega_0^2)}{\omega} \right) \end{aligned} \quad (770)$$

where we have factored out a C and $1/\omega$ from the first expression and used $\omega_0^2 = 1/LC$, the resonance frequency of the circuit.

Resonance for this circuit is the *opposite* of the series LRC circuit we first looked at. It still occurs at the frequency $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$ as before, but now $\frac{1}{Z}$ is *largest* at resonance. To understand how this can be useful, let us think about *current flow* in the circuit both at and away from resonance.

At resonance, the impedance (resistance to current flow) of the L and C together is:

$$\frac{1}{Z_{LC}} = \sqrt{\left(\frac{1}{X_C} - \frac{1}{X_L}\right)^2} = 0 \quad (771)$$

or

$$Z_{LC} = \infty \quad (772)$$

No current flows into the L and C in combination – they behave like an *open circuit* at the resonant frequency. All the current that flows from the voltage at this frequency therefore flows through the resistance (or “load”).

Far from resonance on either side, either χ_C or χ_L will be very *small* – in particular much less than R . The current produced by the voltage will thus find either the capacitor (for high frequencies) or the inductor (for low frequencies) to be a much easier path to ground, provided only that the load resistance R is bigger.

If the voltage were an *ideal* voltage with $r = 0$, capable of delivering *any* amount of current, this wouldn't matter. As the impedance of the parallel LC combination drops, it would simply provide more current and maintain its voltage, while continuing to deliver as much current to the resistor as before. However, many voltage sources – in particular a *radio antenna* – have a significant impedance/resistance of their own, and if they are provided with an easy path to ground this *shorts out* the antenna by pulling enough current from it so that its pole voltage drops to zero (or at any rate a very small number), reducing the current through the resistor to zero at the same time.

This suffices to show that there should be a *maximum* power delivered to the resistance when one is at resonance and the current has no alternative pathway to ground through the LC combination, but it does not suffice to show what the characteristics of the power curve are. To solve this problem exactly, one has to write Kirchoff's laws for the entire circuit, reduce them to an algebraic form, and then solve that form. This is rather painful to do working with trig functions, somewhat easier with complex exponentials, and beyond the scope of this course.

However, we can at least comment on certain aspects of the solution and show a curve or two (for the benefit of any would-be crystal radio builders). First, although it is far from obvious, the power delivered to the load resistor (headphones) will be maximum if its resistance more or less *matches* the resistance of the antenna. This is called “impedance matching” (impedance because in general one has to account for more than just resistance). One can in fact prove a result known as the *Maximum Power Theorem*⁹⁴ 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

$$Z_L = Z_S^* \quad (773)$$

or the impedance of the load has the same amplitude but the opposite phase 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! Hence one needs to design a radio (when possible) to match the impedance of the antenna one hopes to use with it; if one doesn't one either burns too much of the received energy in the antenna itself (when the impedance of the load is too small) or one eliminates one's ability to discriminate the signal.

We can do a somewhat sloppy job of estimating the power delivered to the load resistor with the following argument. Suppose Z is the impedance of the parallel circuit above and r is the resistance of the source. Then we expect the total impedance of the circuit to be $Z' = r + Z$ (where if we don't use complex numbers we will have to separate out and add separately the resistive component of Z to r). The *total* current drawn from the source is thus approximately $I_0 = V_0/Z'$. We can then find the “corrected” source voltage across the resistance R as a (phase shifted) $V_R = V_0 - I_0 r$, and the power delivered to it is thus approximately:

$$P_R = \frac{V_R^2}{R} \quad (774)$$

We plot this very approximate function, computed in just this way, for a range of values of ω around the resonant frequency $\omega_0 = 1/LC = 1.0$ as before in figure 133. The voltage and resistance have been mutually adjusted to make the picture pleasing, with $r = 10\Omega$. Note that we do indeed see peak power delivery to the load when $R \approx 10\Omega$ as expected at least for the three values for R shown. Note how the Q value of the circuit visibly changes with R for the fixed L as well.

⁹⁴Wikipedia: http://www.wikipedia.org/wiki/Maximum_Power_Theorem.

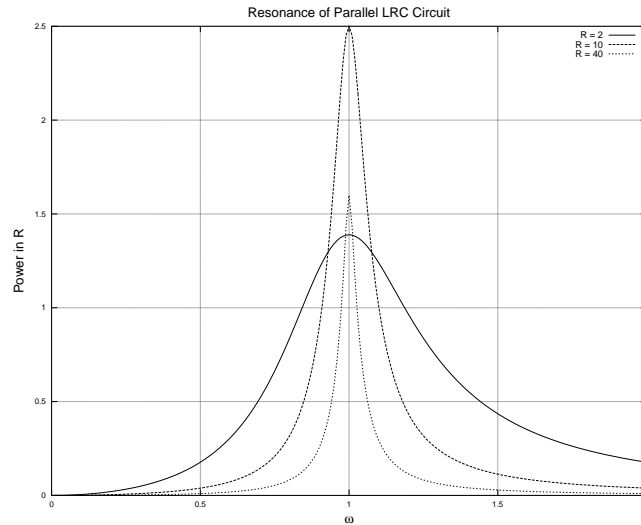


Figure 133: Parallel resonance power delivery in a greatly simplified resistive model.

In the next section we will see how to make practical use of the parallel LRC circuit (and a rectifier) in the design of a *crystal radio*, an inexpensive device capable of receiving, discriminating, and decoding an AM-encoded signal.

The AM Radio and Bandwidth

The simplest 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 134. 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

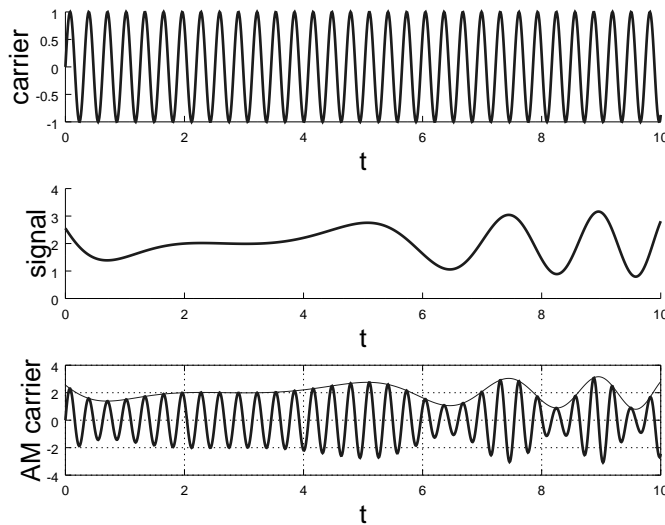


Figure 134: (a) The unencoded carrier with an arbitrary normalization voltage $V_c = 1$ volt and angular frequency ω_0 . (b) The signal to be encoded. A *DC bias* has been added to the AM signal so that the voltage is always positive. This DC bias can be removed at the far end with a simple high-pass 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.

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 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 135, I arbitrarily show a variable C (that’s the arrow symbol) and also introduce the symbol for an antenna and ground. The resistance

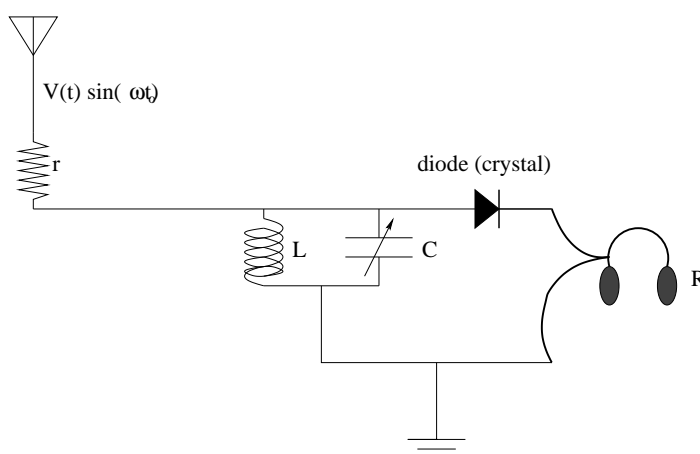


Figure 135: A very simple, idealized crystal radio circuit using a variable capacitor instead of variable inductance (or variable both). Note also the presence of a *diode decoder* – a one-way gate for current (which flows only in the direction of the “arrow”).

r 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 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 semiconducting 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 136 (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.

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

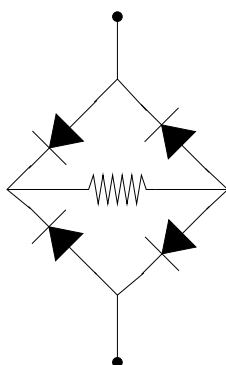


Figure 136: A full-wave rectifier made out of four diodes. The “headphones” are the resistance in the center of the diamond of diodes. Verify that the current always passes through this resistor from left to right, regardless of whether the voltage difference top to bottom is positive or negative.

flipped (but not smeared) signal is shown below in figure 137. 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.

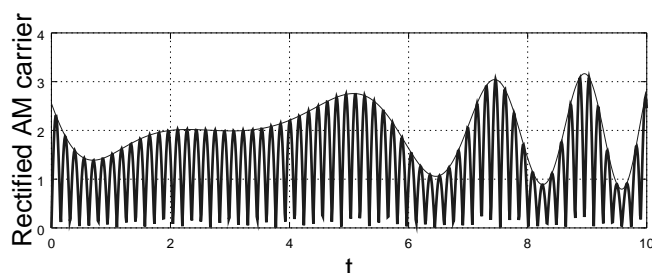


Figure 137: The AM encoded signal after it has been received by a tuned, band-pass filter and full-wave rectified. Note that the *average* output voltage will very closely track the original signal.

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 form *radio* waves, like the AM encoded carrier wave we have just studied. In others, however, those waves are what we know as *light*!

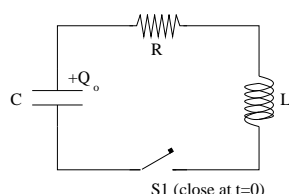
Homework for Week 9

Problem 1.

Physics Concepts

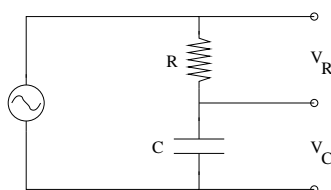
Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



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* $Q(t)$, and draw a qualitatively correct picture of $Q(t)$ in the case that the oscillation is only weakly damped. Show all your work.

Problem 3.



In the circuit above, the AC voltage is $V_0 \cos(\omega t)$. Find:

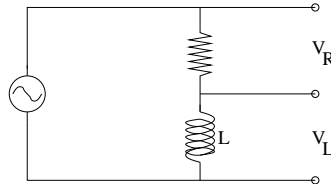
- The current $I(t)$ through the resistor and capacitor, assuming no current is diverted into the branches on the right. Clearly identify the relative phase shift δ between the applied voltage and the current.
- The voltage $V_R(t)$ across the resistor. Factor your answer out so that it is in terms of the dimensionless ωRC .
- The voltage $V_C(t)$ across the capacitor.

This circuit is called a *high-pass filter*, one that delivers the maximum *current* in the circuit only when $\omega RC \gg 1$ (so that the capacitor behaves like a “short” with very low reactance).

When the frequency is *low*, the capacitor acts like a gap, with very high reactance, and does not permit current to flow. At this point the applied voltage drop across the capacitor is maximal, and this pair of tap points is sometimes used to help clean up a DC power supply by “shorting out”

high frequency pulses while maintaining a steady *DC* voltage across the fully charged capacitor. In this configuration, the capacitor can also serve as a reservoir of charge and can maintain the voltage even if the load imposes a transient peak in demand that is higher than the supply voltage source could otherwise handle.

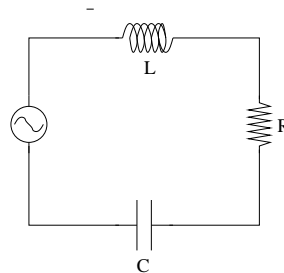
Problem 4.



Repeat the previous problem for the *LR* circuit above, evaluating $I(t)$, δ , $V_R(t)$, $V_L(t)$ in terms of the dimensionless $\frac{\omega L}{R}$. This circuit is used as a *low pass filter*, with peak current through and voltage across R at *low* frequencies, while high frequencies are blocked by the inductor.

When might one wish to use the V_L versus the V_R voltage taps, respectively? Think about this: Not all loads are *resistive*...

Problem 5.



A *series* LRC circuit connected across a variable AC voltage source $V = V_0 \cos(\omega t)$ is drawn above. Find:

- The current $I(t)$ in the primary supply wire (as shown in the figure above) with all terms, e.g. the phase δ , and the impedance Z defined (the latter in terms of the individual reactances).
- The *average* power dissipated by the circuit. Remember, $P(t) = V(t)I(t)$ (where $V(t)$ is the voltage across each circuit element and $I(t)$ is the common current through it). Two of the circuit elements have *zero* average power, but you must prove this.

Hint: To find the answer you must assume that:

$$I(t) = I_0 \cos(\omega t - \delta)$$

and then add the *voltage drop* across each series element. Your answer for Z should remind you of series addition of resistors, using reactances instead (two of which are $\pi/2$ out of phase with the resistor).

Problem 6.

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 did part b) of the previous problem correctly, you should have found that:

$$P_{\text{av}}(\omega) = I_{\text{av}}^2 R = \frac{V_{\text{rms}}^2 R \omega^2}{L^2(\omega^2 - \omega_0^2)^2 + \omega^2 R^2}$$

is the average power delivered to the circuit by the voltage and is also the average power “burned” by the resistor, since 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.

Show that for a sharply peaked resonance (one with large Q):

$$\Delta\omega \approx \frac{R}{L}$$

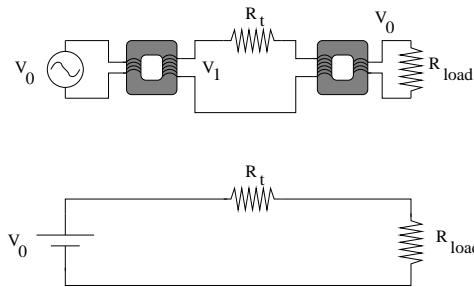
so that

$$Q = \frac{\omega_0}{\Delta\omega} \approx \frac{\omega_0 L}{R}$$

where $\Delta\omega$ is the full width at half maximum of the power curve you derive in the first part.

To do this, set the expression above equal to the *computed* half-maximum power, and solve for the two quadratic roots for ω , assuming that both of them are *very close* to (but not equal to) ω_0 (this is the sharply peaked part). You may find the following factorization useful:

$$\omega^2 - \omega_0^2 = (\omega - \omega_0)(\omega + \omega_0)$$

Problem 7.


In this problem you must analyze the problem of power transmission that dominated the famous Edison vs Tesla “war” that took place some hundred years ago. Above you can see two alternatives for transmitting power long distances. The first circuit is Tesla’s – generate AC power at a relatively low voltage V_0 (which is easy). Step the power up to a very high voltage $V_1 \gg V_0$ and transmit it at high voltage across a long transmission wire of fixed resistance R_t . Step it back down to voltage V_0 and then place the load R_{load} across it.

The second circuit is Edison’s. Generate a DC voltage V_0 . Transmit it down identical transmission lines and place it across an identical load.

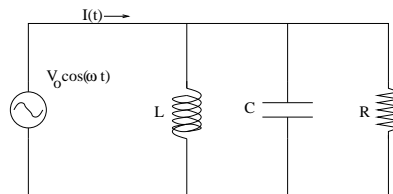
Your job is to *compute the way the power is divided up* between P_{load} (which is fixed – the power we need to light a light bulb, for example) and P_t , the power wasted heating up the transmission lines. The better solution has $P_t \ll P_{\text{load}}$. Find a relationship between the ratios:

$$\frac{V_0}{V_1}$$

and

$$\frac{P_t}{P_{\text{load}}}$$

that proves that Tesla’s solution wins (and by how much it wins, given “reasonable” estimates for R_t/R_{load}).

Problem 8.


A *parallel* LRC circuit connected across a variable AC voltage source $V = V_0 \cos(\omega t)$ is drawn above. Find:

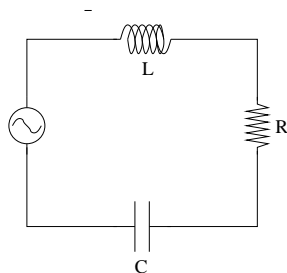
a) The current $I(t)$ in the primary supply wire (as shown in the figure above) with all terms, e.g. the phase δ , and the impedance Z defined (the latter in terms of the individual reactances).

b) The *average* power dissipated by the circuit. Note that (if you are clever and remember what each element does in the circuit) you don’t really *have* to solve a) to get this answer, although you can certainly get the same answer from a knowledge of $V(t)$ and $I(t)$ and some integration.

Hint: To find the answer you must add the *currents* being drawn by each element separately!

Your answer for Z should remind you of parallel addition of resistors, using reactances instead (two of which are $\pi/2$ out of phase with the resistor, of course).

Advanced Problem 9.



This problem is in two parts. First, *for your own enduring benefit* I want *you* to derive the full solution to the driven LRC circuit problem. In particular, start with Kirchhoff's rule for the loop and either assume a complex $V(t) = V_0 e^{i\omega t}$ and $I(t) = I_0 e^{i\omega t}$ (where by convention V_0 is real, $I_0 = |I_0| e^{-i\delta}$, and where one gets physical answers at the *end* by taking the real part of the complex answers, *or* assume $V(t) = V_0 \cos(\omega t)$ and $I(t) = I_0 \cos(\omega t - \delta)$. Find an algebraic expression that expresses the sum of the voltages. Solve this expression using either phasors (which will work in both cases, one in the complex plane and one in a "real" x-y plane) or in the complex case directly using algebra, no pictures really required.

Factor out the solution to obtain $|I_0|$ and δ , Z (the impedance), and the voltages across each element as a function of time.

Week 10: 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*:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \left(\int_{S/C} \vec{J} \cdot \hat{n} dA + \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{E} \cdot \hat{n} dA \right)$$

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*:

$$I_{MDC} = \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{E} \cdot \hat{n} dA$$

- It is worth writing down the complete set of trading cards, suitable for engraving:

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (775)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \quad (776)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \left(\int_{S/C} \vec{J} \cdot \hat{n} dA + \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{E} \cdot \hat{n} dA \right) \quad (777)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (778)$$

- 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):

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (779)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} - \frac{d}{dt} \mu_0 \epsilon_0 \int_{S/C} \vec{E} \cdot \hat{n} dA = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \quad (780)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = 0 \quad (781)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} + \frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA = 0 \quad (782)$$

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:

$$\frac{\partial^2 \vec{E}}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \quad (783)$$

$$\frac{\partial^2 \vec{B}}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \vec{B}}{\partial t^2} = 0 \quad (784)$$

The $\frac{\partial^2}{\partial z^2}$ symbol in this expression, let me remind you, just means to take the derivative of the functions $\vec{E}(\vec{x}, t)$ and $\vec{B}(\vec{x}, t)$ with respect to the z -coordinate only, pretending that the other coordinates are constants. In this equation,

$$c = \sqrt{\frac{k_e}{k_m}} = \frac{1}{\sqrt{\epsilon_0 \mu_0}} = 3 \times 10^8 \text{ meters per second} \quad (785)$$

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:

$$E_x(z, t) = E_{0x} \sin(kz - \omega t) \quad (786)$$

$$B_y(z, t) = B_{0y} \sin(kz - \omega t) \quad (787)$$

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*:

$$E_y(z, t) = E_{0y} \sin(kz - \omega t) \quad (788)$$

$$B_x(z, t) = -B_{0x} \sin(kz - \omega t) \quad (789)$$

that also propagate in the z -direction, as determined from the derivation of the wave equations above.

In these equations, note well that:

$$k = \frac{2\pi}{\lambda} \quad (790)$$

is the *wave number* of the wave, where λ is the *wavelength* of the harmonic wave, while:

$$\omega = \frac{2\pi}{T} \quad (791)$$

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:

$$c = \frac{\lambda}{T} = \frac{\omega}{k} \quad (792)$$

for a harmonic wave.

- The flow of energy in an electromagnetic wave (and field in general) can be determined from the *Poynting vector*:

$$\vec{S} = \frac{1}{\mu_0}(\vec{E} \times \vec{B}) \quad (793)$$

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:

$$I = \frac{dP}{dA} = \frac{d}{dA} \frac{dU}{dt} = |S| \quad (794)$$

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*:

$$P_A = \int_A \vec{S} \cdot \hat{n} \, dA \quad (795)$$

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{E} or \vec{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:

$$\Delta U = \left(\frac{1}{2} \epsilon_0 E_x^2 + \frac{1}{2\mu_0} B_y^2 \right) A c \Delta t \quad (796)$$

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:

$$\frac{\Delta U}{A \Delta t} = \frac{1}{\mu_0} |\vec{E}_x| |\vec{B}_y| \quad (797)$$

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:

$$|p_f| = \frac{U}{c} \quad (798)$$

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:

$$P_r = \frac{I_{\text{thru } A}}{c} \quad (799)$$

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:

$$F_S = \frac{1}{c} \int_A \vec{S} \cdot \hat{n} \, dA \quad (800)$$

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:

$$F_S = \frac{SA}{c} \quad (801)$$

or

$$F_S = \frac{2SA}{c} \quad (802)$$

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:

$$\frac{dP}{d\Omega} = \frac{q^2}{16\pi^2\epsilon_0} \frac{1}{c^3} \left| \frac{d^2z}{dt^2} \right|^2 \sin^2(\theta) \quad (803)$$

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 must 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:

$$\frac{dP}{d\Omega} = \frac{c^2}{32\pi^2} \sqrt{\frac{\mu_0}{\epsilon_0}} k^4 |p_z|^2 \sin^2(\theta) \quad (804)$$

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:

$$I(\theta) = \frac{P_0}{r^2} \sin^2(\theta) \quad (805)$$

(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:

$$P_0 = I_{\text{rms}}^2 R_{\text{rad}} \quad (806)$$

where I_{rms} is the root-mean-square current in the antenna and R_{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.

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:

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (807)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \quad (808)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \int_{S/C} \vec{J} \cdot \hat{n} dA \quad (809)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (810)$$

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:

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (811)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \int_{S/C} \vec{J}_e \cdot \hat{n} dA \quad (812)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = 0 \quad (813)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} + \frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA = 0 \quad (814)$$

I put a tiny $_e$ subscript on the \vec{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*⁹⁵ but that big hole is *ugly*, as is (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 ultimately nature itself is 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. Experimentalists today search for magnetic monopoles; we ourselves will follow in Maxwell's footsteps and search for the missing term.

⁹⁵Seriously. 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...

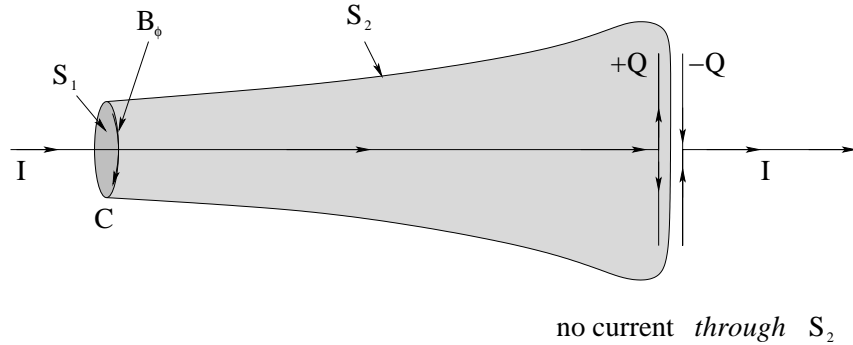


Figure 138: A simple circuit and pair of surfaces that illustrate how Ampere's Law is (so far) *wrong*, with two completely different currents for the two surfaces S_1 and S_2 .

Fortunately, we've learned enough at this point to be able to see that Ampere's Law is *obviously wrong*! Consider the following specific example. In figure 138 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:

$$\oint_C \vec{B} \cdot d\vec{\ell} = B_\phi 2\pi r = \mu_0 \int_{S_1/C} \vec{J} \cdot \hat{n} dA = \mu_0 I \quad (815)$$

(where recall that S/C should be read as “the open surface S bounded by the closed curve C ”) so that

$$B_\phi = \frac{\mu_0 I}{2\pi r} \quad (816)$$

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 a plane). However, there is nothing in the *mathematics* of Ampere's Law that requires me to use that particular surface.

I could choose to use 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:

$$I = \int_{S_2/C} \vec{J} \cdot \hat{n} dA = 0(!) \quad (817)$$

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.

One way we can try to deal with this is to insist that we have to use “nice” curves C (ones in a plane, for example) and “nice” surfaces S (ones in that same plane, for example) but that isn’t very satisfactory – it seems like just a way of saying that Ampere’s Law is really just Ampere’s Sort of OK Rule That Sometimes Works, Sometimes, If We Cheat. We want a natural law to *always* work – it has to be “unbreakable”, especially by as simple a thing as bending C into twisted loop (like a crumpled coat hanger) or choosing “the wrong” S (by what standard? how can we decide it is wrong without knowing the answer some other way?).

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)⁹⁶.

Our job, it appears, is to try to make the current in Ampere’s Law invariant so that it gives us the exact same current for any surface S/C we might *happen* to choose to solve a problem. That way we’ll all get the same answer for B_ϕ , and if we choose the *right* invariant current (there may be more than one) that answer will even agree with experiment!

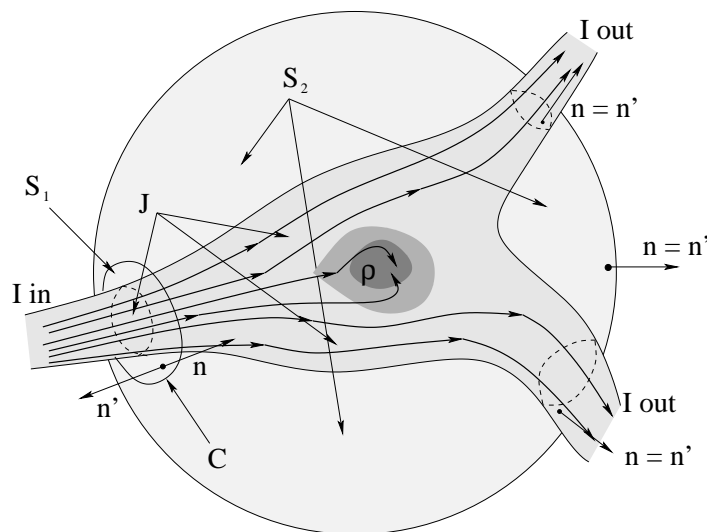


Figure 139: A very general current density flows through space. Some current flows in from the left and exits on the right, but some builds up in the current density ρ in the volume between the two surfaces S_1 and S_2 . The point is that the *difference* between the flux (current) in through S_1 and 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 139. 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*. We will discover that Ampere’s Law fails to account for charge conservation *and* Gauss’s Law for Electricity consistently. Even better, 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.

⁹⁶In 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-relativistic physics* – and was backed up by numerous experiments that showed that he was *right* to do so! Kind of scary, that...

Note that 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* surface, one that encloses a volume V (inside the sphere). The current density \vec{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. If charge is conserved, 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:

$$\int_{S_1/C} \vec{J} \cdot \hat{n} dA - \int_{S_2/C} \vec{J} \cdot \hat{n} dA = \frac{d}{dt} \oint_{V/S} \rho dV \quad (818)$$

In this equation, the normals \hat{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. Note that we first derived/discussed this law as equation 300 way back in the week where we first derived and discussed current and resistance and defined current density in the first place, so this *shouldn't* be a complete mystery even if you've forgotten the first pass through it.

The integral on the left looks strangely familiar. In fact, it is part of Gauss's Law for Electricity! Using Gauss's Law we can *reexpress* it:

$$\oint_{V/S} \rho dV = \epsilon_0 \oint_S \vec{E} \cdot \hat{n}' dA \quad (819)$$

Now we'll break this up into two pieces – we can surely integrate this over S_1 and S_2 separately, as long as $S_1 + S_2 = S$. However, I'm going to make one more change. The normal \hat{n}' in the Gauss's Law expression is (recall) the *outward directed normal*. This goes from left to right on S_2 , but on S_1 it goes from right to left! I want to make \hat{n} exactly the same as in the integrals on the left hand side of my expression of charge conservation, so I have to change the sign of the S_1 integral:

$$\begin{aligned} \oint_{V/S} \rho dV &= \epsilon_0 \oint_S \vec{E} \cdot \hat{n}' dA \\ &= -\epsilon_0 \int_{S_1/C} \vec{E} \cdot \hat{n} dA + \epsilon_0 \int_{S_2/C} \vec{E} \cdot \hat{n} dA \end{aligned} \quad (820)$$

Now we substitute this back into our original equation to get:

$$\int_{S_1/C} \vec{J} \cdot \hat{n} dA - \int_{S_2/C} \vec{J} \cdot \hat{n} dA = -\epsilon_0 \frac{d}{dt} \int_{S_1/C} \vec{E} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{E} \cdot \hat{n} dA \quad (821)$$

Last, we move all of the S_1 integrals to the left, and all of the S_2 integrals to the right:

$$\int_{S_1/C} \vec{J} \cdot \hat{n} dA \epsilon_0 \frac{d}{dt} + \int_{S_1/C} \vec{E} \cdot \hat{n} dA = \int_{S_2/C} \vec{J} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{E} \cdot \hat{n} dA \quad (822)$$

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:

$$I_{\text{invariant, through } C} = \int_{S/C} \vec{J} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA \quad (823)$$

where the result now holds for *any* surface S bounded by any give closed curve C !

Let us now guess that this invariant current is the *correct* one to use in Ampere's Law, and see if it gives us the right answer in at least one problem where we *know* the answer and Ampere's Law as it was before got it wrong. That is, suppose Ampere's Law is really:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \left\{ \int_{S/C} \vec{J} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA \right\} \quad (824)$$

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 use this expression to compute I in our capacitor problem above, when we compute the *invariant* current through S_1 we still get I (because the field due to the capacitor is confined to live in between the plates of the capacitor and doesn't pass through S_1 . If we apply it to the surface S_2 , no physical current gets through, but the *field inside the capacitor* is (recall):

$$E = \frac{Q}{\epsilon_0 A} \quad (825)$$

where A is the area of the capacitor. The integral:

$$\int_{S_2/C} \vec{E} \cdot \hat{n} dA = EA = \frac{Q}{\epsilon_0} \quad (826)$$

and hence

$$I_{\text{inv}} = \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{E} \cdot \hat{n} dA = EA = \frac{dQ}{dt} = I \quad (827)$$

because I is the rate at which the capacitor is charging! *We get the same I for both surface!* We therefore get the same (correct) magnetic field around C from both surfaces.

The extra term we have added to the physical current was originally added by James Clerk Maxwell, and the implications of this term were 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*:

$$I_{\text{MDC}} = \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{E} \cdot \hat{n} dA \quad (828)$$

From now on we will assume that equation 824 is the actual, correct form for Ampere's Law, the one that will always give the right answer, the law of nature. 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 is varying.

In fact, there is one very important place where it fails. It fails to describe the magnetic field *inside* the parallel plate capacitor. Let's work that out as an example.

Example 10.0.1: The Magnetic Field Inside a Parallel Plate Capacitor

In figure 140 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).

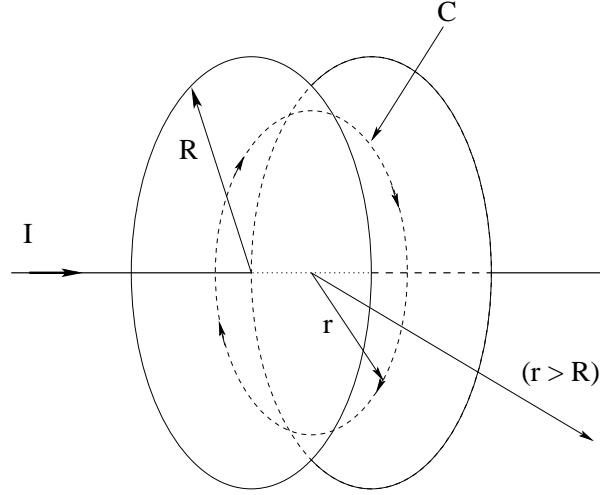


Figure 140: A capacitor made up of two circular disks is being charged by a current I . The increasing electric field between the two plates becomes a *Maxwell Displacement Current* that creates a magnetic field identical to the one that would exist inside a uniform conductor of the same radius (assuming the conductor had a magnetic permeability and electric permittivity identical to the vacuum value, not really a very good assumption).

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):

$$E = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 A} \quad r < R \quad (829)$$

(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.

Second, we have to evaluate the flux through the Amperian path C (for $r < R$) in figure 140:

$$\phi_C = \int_{S/C} \vec{E} \cdot \hat{n} dA = EA = \frac{Q\pi r^2}{\epsilon_0 A} = \frac{Q\pi r^2}{\epsilon_0 \pi R^2} \quad (830)$$

(where we have used $A = \pi R^2$ at the end).

Third, we have to write Ampere's Law for this Amperian path:

$$\oint_C \vec{B} \cdot d\vec{\ell} = B_\phi 2\pi r = \mu_0 \left(\int_{S/C} \vec{J} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA \right) \quad (831)$$

$\vec{J} = 0$ (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:

$$B_\phi 2\pi r = \mu_0 \frac{dQ}{dt} \frac{r^2}{R^2} = \frac{\mu_0 I r^2}{R^2} \quad (832)$$

We rearrange this to obtain half of our answer:

$$B_\phi = \frac{\mu_0 I r}{2\pi R^2} \quad r < R \quad (833)$$

(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:

$$B_\phi = \frac{\mu_0 I}{2\pi r} \quad r > R \quad (834)$$

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.

10.1: 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:

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (835)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \quad (836)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \left(\int_{S/C} \vec{J} \cdot \hat{n} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA \right) \quad (837)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (838)$$

The *symmetry* will now be 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:

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \quad (839)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} - \mu_0 \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA = \mu_0 \int_{S/C} \vec{J}_e \cdot \hat{n} dA \quad (840)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = 0 \quad (841)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} + \frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA = 0 \quad (842)$$

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 $\vec{\nabla}$ 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.

We start by doing away with one complication – the sources. Note that *ultimately* both electric and magnetic fields have to come from *electric charges* – the only in Maxwell's Equations that get electric or magnetic fields into the Universe in the first place are those pesky little charges, but again, to understand how they make them *correctly* we really need to lose the integrals and work with differential equations and we're not ready to do that yet (and never will be, in this course). So here is a very short discursion on sources of electromagnetic fields:

10.1.1: Accelerating Charge

Against my custom I'm not deriving *anything* in this short section. Either you believe me or you don't, or you read a book or take an advanced course that does it right⁹⁷. Just be sure that you take two or three fairly serious courses in ordinary and partial differential calculus and maybe complex variables first...

If one takes an electric charge and accelerates it, it radiates away electric and magnetic energy in the form of electromagnetic waves of the sort we're about to derive. 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).

That's it. Not complicated at all (although the derivation of this fact is a bit hairy).

Well, when do charges accelerate? Well, they'd accelerate if they (for example) went around in a circular orbit. That pesky centripetal acceleration qualifies as one that would radiate energy. They'd also accelerate if they were just oscillating harmonically, as a harmonic oscillator in one dimension is nearly always accelerating.

These two observations are among the most profound in all of physics. What they add up to is this: *There is no obvious way to make a model for an atom that does not involve orbiting, 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 an atom consisting of a proton being orbited by a light electron, you find that it collapses, with the electron spiralling into the proton while it radiates away energy, in around 10^{-20} seconds. A classical Universe based on Maxwell's equations would last just about that long.

⁹⁷Such as my grown-up graduate E&M book online, used in a graduate course in Classical Electrodynamics. Even most undergrad intermediate E&M courses do a sloppy job of treating radiation from sources, partly because the math required is relatively *difficult*

Either Maxwell's Equations are *wrong*, or *classical Newtonian mechanics itself* is wrong!

In which case everything we've learned over the last two semesters is wrong.

Too bad, ladies and gentlemen. Maxwell's Equations appear to be *correct*. Classical Mechanics is *not*. Visualize Newton, spiralling down to the earth in flames, never to rise again as high as he was before Maxwell made this momentous discovery. Over the last century it has been replaced by *quantum* mechanics, a *wave* theory of matter that is, to say the least, a lot more complicated than the relatively *simple* $\vec{F} = m\vec{a}$ that has governed nearly everything so far.

The last thing I want to mention before we return to our regularly scheduled (but false) "classical" treatment of electromagnetism is that *nearly all* electromagnetic radiation comes from oscillating *dipoles*, predominantly electric dipoles at that. Our Lorentz model atom turns out to be very useful (all the way into graduate classical mechanics) for understanding the "generic" properties of radiating atoms and matter interacting with a time dependent electromagnetic field.

Back to work.

10.1.2: The Wave Equation

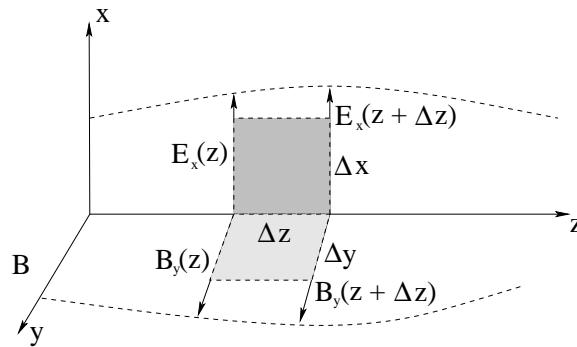


Figure 141: Two particular components of the electric and magnetic field, in a coordinate frame "far" from any sources and varying in space and time. The graph is a snapshot at a particular time t , but we can imagine that $E_x(z, t)$ and $B_y(z, t)$ generally and ignore any other variation with x or 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:

$$\oint_S \vec{E} \cdot \hat{n} dA = 0 \quad (843)$$

$$\oint_S \vec{B} \cdot \hat{n} dA = 0 \quad (844)$$

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{E} \cdot \hat{n} dA \quad (845)$$

$$\oint_C \vec{E} \cdot d\vec{\ell} = -\frac{d}{dt} \int_{S/C} \vec{B} \cdot \hat{n} dA \quad (846)$$

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{E} and \vec{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 141, 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 variation⁹⁸. 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{B}). That means that we need to go around the loop *counterclockwise* as drawn in the page. Then:

$$\begin{aligned} \oint \vec{E} \cdot d\vec{\ell} &= -\frac{d}{dt} \int_{\Delta A} \vec{B} \cdot \hat{n} dA \\ 0 \cdot \Delta z + E_x(z + \Delta z) \Delta x - 0 \cdot \Delta z - E_x(z) \Delta x &= -\frac{d}{dt} (B_y \Delta A) \\ (E_x(z + \Delta z) - E_x(z)) \Delta x &= -\frac{dB_y}{dt} \Delta x \Delta z \\ \frac{(E_x(z + \Delta z) - E_x(z))}{\Delta z} &= -\frac{dB_y}{dt} \end{aligned} \tag{847}$$

where we do the loop piecewise and get no contribution when we go in the z direction (because \vec{E} is in the x -direction perpendicular to z). If we take the limit $\Delta z \rightarrow 0$ of the left hand side this is just the *definition* of the derivative and we get⁹⁹:

$$\frac{dE_x}{dz} = -\frac{dB_y}{dt}$$

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{E} in the x direction, or again counterclockwise as seen on the page from above. The *only* term on the right is the

⁹⁸It 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.

⁹⁹Technically, 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.

Maxwell Displacement Current – this is where Maxwell's contribution shines!

$$\begin{aligned}
 \oint \vec{B} \cdot d\vec{\ell} &= \mu_0 \epsilon_0 \frac{d}{dt} \int_{\Delta A} \vec{E} \cdot \hat{n} dA \\
 B_y \Delta y + 0 \cdot \Delta z - B_y(z + \Delta z) \Delta y - 0 \cdot \Delta z &= \mu_0 \epsilon_0 \frac{d}{dt} (E_x \Delta A) \\
 - (B_y(z + \Delta z) - B_y(z)) \Delta y &= \mu_0 \epsilon_0 \frac{dE_x}{dt} \Delta y \Delta z \\
 \frac{(B_y(z + \Delta z) - B_y(z))}{\Delta z} &= -\mu_0 \epsilon_0 \frac{dE_x}{dt} \\
 \frac{dB_y}{dz} &= -\mu_0 \epsilon_0 \frac{dE_x}{dt}
 \end{aligned}$$

where we have taken the limit $\Delta z \rightarrow 0$ as before in the last step¹⁰⁰.

Since we're going to use these two results a *lot*, let's write them down right next to each other:

$$\frac{dE_x}{dz} = -\frac{dB_y}{dt} \quad (848)$$

$$\frac{dB_y}{dz} = -\mu_0 \epsilon_0 \frac{dE_x}{dt} \quad (849)$$

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 141). In that case one expects to get:

$$\frac{dE_y}{dz} = \frac{dB_x}{dt} \quad (850)$$

$$\frac{dB_x}{dz} = \mu_0 \epsilon_0 \frac{dE_y}{dt} \quad (851)$$

from an identical argument to the one above, something you can verify by completely recapitulating the derivation above as part of your homework¹⁰¹.

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 :

$$\frac{d}{dz} \frac{dE_x}{dz} = \frac{d^2 E_x}{dz^2} = -\frac{d}{dz} \frac{dB_y}{dt} = -\frac{d}{dt} \frac{dB_y}{dz} \quad (852)$$

¹⁰⁰Once 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.

¹⁰¹Sure, sure, they should all be partials. In fact, you are basically deriving:

$$\begin{aligned}
 \vec{\nabla} \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} \\
 \vec{\nabla} \times \vec{B} &= \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t},
 \end{aligned}$$

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 independent 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...

If we substitute the second equation in for the last term, we get:

$$\frac{d^2 E_x}{dz^2} = -\frac{d}{dt} \mu_0 \epsilon_0 \frac{dE_x}{dt} = \mu_0 \epsilon_0 \frac{d^2 E_x}{dt^2} \quad (853)$$

or

$$\frac{d^2 E_x}{dz^2} - \mu_0 \epsilon_0 \frac{d^2 E_x}{dt^2} = 0 \quad (854)$$

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*:

$$\frac{d^2 y(x, t)}{dx^2} - \frac{1}{v^2} \frac{d^2 y(x, t)}{dt^2} = 0 \quad (855)$$

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:

$$v^2 = \frac{1}{\mu_0 \epsilon_0} = \frac{4\pi}{\mu_0 4\pi \epsilon_0} = \frac{4\pi}{\mu_0} \frac{1}{4\pi \epsilon_0} = \frac{k_e}{k_m} \quad (856)$$

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:

$$v^2 = \frac{9 \times 10^9}{10^{-7}} = 9 \times 10^{16} \frac{\text{meters}^2}{\text{second}^2} \quad (857)$$

or:

$$v = c = 3 \times 10^8 \frac{\text{meters}}{\text{second}}. \quad (858)$$

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 electro-magnetic 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 interaction¹⁰²?

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:

$$\frac{d^2 B_y}{dz^2} - \mu_0 \epsilon_0 \frac{d^2 B_y}{dt^2} = \frac{d^2 B_y}{dz^2} - \frac{1}{c^2} \frac{d^2 B_y}{dt^2} = 0 \quad (859)$$

¹⁰²The correct answer: not much...

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 wave¹⁰³.

10.2: 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 part of the course. Recall that we showed that a very general solution to the wave equation for waves on a string was:

$$y(x, t) = f(x \pm vt) \quad (860)$$

where $f(u)$ is an arbitrary one-dimensional function. Basically *any* functional form that propagates to the right or left along the x -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. We will thus *ignore* this *general* 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, 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, we'd expect the field to have the form of a *harmonic wave*:

$$E_x(z, t) = E_{0x} \sin(kz \pm \omega t) \quad (861)$$

¹⁰³Even 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?

where ω is the frequency of the oscillating dipole source that is producing the wave¹⁰⁴. We are fortunate in this is actually a function of the form $f(z \pm vt)$! To see this, let's factor the argument:

$$E_x(z, t) = E_{0x} \sin(k(z \pm \frac{\omega}{k}t)) = E_{0x} \sin(k(z \pm ct)) \quad (862)$$

which has the desired form if $c = \omega/k$. Indeed, if you substitute this harmonic wave into the wave equation, you get:

$$\begin{aligned} \frac{d^2}{dz^2} E_{0x} \sin(kz \pm \omega t) &= -k^2 E_{0x} \sin(kz \pm \omega t) &= \frac{1}{c^2} \frac{d^2}{dt^2} E_{0x} \sin(kz \pm \omega t) \\ &= -\frac{1}{c^2} \omega^2 E_{0x} \sin(kz \pm \omega t) \end{aligned} \quad (863)$$

or (dividing out)

$$c^2 = \frac{\omega^2}{k^2} \quad (864)$$

and $c = \omega/k$ as promised.

Again recalling our work with harmonic waves, we expect that in these equations:

$$k = \frac{2\pi}{\lambda} \quad (865)$$

is the *wave number* of the wave, the “spatial angular frequency” in terms of the *wavelength* of the wave λ , just as:

$$\omega = \frac{2\pi}{T} \quad (866)$$

is the *temporal* angular frequency of the wave in terms of its period T . Thus:

$$c = \frac{\omega}{k} = \frac{2\pi}{T} \frac{\lambda}{2\pi} = \frac{\lambda}{T} = f\lambda \quad (867)$$

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.

We expect that:

$$B_y(z, t) = B_{0y} \sin(kz \pm \omega t + \phi) \quad (868)$$

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*:

$$E_x(z, t) = E_{0x} \sin(kz - \omega t) \quad (869)$$

$$B_y(z, t) = B_{0y} \sin(kz - \omega t + \phi) \quad (870)$$

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:

$$\frac{d}{dz} E_{0x} \sin(kz - \omega t) = -\frac{d}{dt} B_{0y} \sin(kz - \omega t + \phi) \quad (871)$$

$$k E_{0x} \sin(kz - \omega t) = \omega B_{0y} \sin(kz - \omega t + \phi) \quad (872)$$

$$E_{0x} \sin(kz - \omega t) = \frac{\omega}{k} B_{0y} \sin(kz - \omega t + \phi) \quad (873)$$

$$E_{0x} \sin(kz - \omega t) = c B_{0y} \sin(kz - \omega t + \phi) \quad (874)$$

¹⁰⁴Note 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.

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:

$$E_{0x} = cB_{0y} \quad (875)$$

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:

$$E_y(z, t) = E_{0y} \sin(kz - \omega t) \quad (876)$$

$$B_x(z, t) = B_{0x} \sin(kz - \omega t + \phi) \quad (877)$$

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:

$$\frac{d}{dz} E_{0y} \sin(kz - \omega t) = \frac{d}{dt} B_{0x} \sin(kz - \omega t + \phi) \quad (878)$$

$$k E_{0y} \sin(kz - \omega t) = -\omega B_{0x} \sin(kz - \omega t + \phi) \quad (879)$$

$$E_{0y} \sin(kz - \omega t) = -\frac{\omega}{k} B_{0x} \sin(kz - \omega t + \phi) \quad (880)$$

$$E_{0y} \sin(kz - \omega t) = -c B_{0x} \sin(kz - \omega t + \phi) \quad (881)$$

This time we see that the two fields must be in phase and that:

$$E_{0y} = -c B_{0x} \quad (882)$$

For a wave propagating to the right, *both* of the independent components of \vec{E} are related to the coupled components of \vec{B} such that:

$$|vE| = c|\vec{B}| \quad (883)$$

and so that the E-field *crossed into* the B-field points in the direction of the wave's propagation. That is, if we let the fingers of our right hand line up with \vec{E} and curl them into \vec{B} , our thumb points in the direction of propagation. This also works for waves propagating in the $-x$ direction, e.g. $E_{0x} \sin(kz + \omega t)$ (try it!).

10.3: 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:

$$\eta_e = \frac{1}{2} \epsilon_0 E^2 \quad (884)$$

$$\eta_m = \frac{1}{2\mu_0} B^2 \quad (885)$$

Now, 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 spherically symmetric and produced no field – and then the field *reaches* it some time after the dipole is excited and starts to oscillate.

No energy in that region of space before, yes energy after, 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:

$$\Delta U = \frac{1}{2} \left(\epsilon_0 E^2 + \frac{1}{\mu_0} B^2 \right) c \Delta t A \quad (886)$$

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:

$$I = \frac{P}{A} = \frac{\Delta U}{A \Delta t} = \frac{1}{2} \left(\epsilon_0 E^2 + \frac{1}{\mu_0} B^2 \right) c \quad (887)$$

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 $B_y(z, t)$. Then $E_x^2 = E_x(cB_y)$ and $B_y^2 = B_y(E_x/c)$, so if we multiply this out we get:

$$I = \frac{1}{2} \left(\epsilon_0 E_x B_y c^2 + \frac{1}{\mu_0} E_x B_y \right) \quad (888)$$

But $c^2 = \frac{1}{\epsilon_0 \mu_0}$ so that:

$$I = \frac{1}{\mu_0} E_x B_y \quad (889)$$

Note as well, that *again* by a hopefully familiar argument, we derived the above 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 $P = I\Delta A' \cos(\theta)$ or more generally, if we define the “vector intensity” in the direction by:

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \quad (890)$$

– a quantity eponymously named the *Poynting vector* (yes, it *poyn*ts in the direction that the wave propagates, har har) then the power through any surface S is *flux of the Poynting vector* through that surface:

$$P = \int_S \vec{S} \cdot \hat{n} dA \quad (891)$$

Again, I'm hoping that I don't have to do much more than this – sketch out one more example of how the flow of a vector field *through* a surface is conserved and correctly accounted for by the flux integral.

The intensity is thus the *magnitude of the Poynting vector*:

$$I = |\vec{S}| \quad (892)$$

and is still a very useful quantity in its own right.

The Poynting vector is actually pretty much magical. For example, it doesn't just work with dynamic electromagnetic waves – it works for *static fields* as well. In fact, for your homework you will prove that the flux of the Poynting vector into a resistor, and inductor and a capacitor all precisely equal $VI - I^2 R$, $LI dI/dt$ and QI/C respectively. This seems

to suggest that the power that appears as heat in a resistor is actually *electromagnetic energy that flows in through the sides of the resistor*, quite contrary to at least *my* naive expectations. But it gets the answers we obtained other ways precisely correct – it is difficult to argue with the conclusion.

The electromagnetic field doesn't just carry energy – it carries *momentum*¹⁰⁵. If you recall our arguments way back when we discussed the failure of Newton's Third Law, we knew even then that it must be so – the missing momentum has to go *someplace* or momentum violation would be ubiquitous in electromagnetism – but now we have to run it down.

This is actually rather tricky. It isn't easy to derive the momentum carried by the electromagnetic field, because it has no mass. The easiest way to see what it must be is to examine the net force exerted on a point charge in an electromagnetic field. We'll do this (and define the associated *radiation pressure* in the next section).

10.4: Radiation Pressure and Momentum

There are two arguments that make it comparatively simple to see that an electromagnetic wave must exert a force on charged matter that it strikes at a surface. Let's take the simplest one first – an electromagnetic wave incident on the surface of a perfect conductor at right angles.

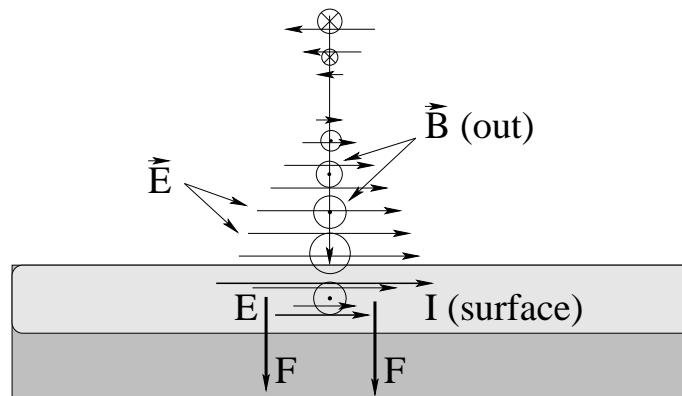


Figure 142: An electromagnetic wave incident on a conducting surface penetrates a short distance into the conductor, inducing a **surface current** in the direction of the electric field at the surface.

Although it is beyond the scope of this course to treat waves incident on conductors, it is a True Fact(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 that cancels the field on the interior that is a *few atoms thick*, time varying fields penetrate a small distance into a conductor (called the **skin depth**) before being cancelled by a 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).

In figure 142 an electromagnetic wave is incident at right angles on a conducting surface. The 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**

¹⁰⁵And often angular momentum as well, but that is beyond the scope of this course.

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). The **magnetic field** also penetrates a short distance into the surface and exerts a force on this surface current. As you can see from the figure, this force is expected to be **in the direction of the wave** 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), it carries *linear momentum* (per unit time) and exerts a force on any conducting surface it collides with. 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.

Indeed, the 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.

We will idealize these two rules and assume that absorption transfers exactly the momentum of wave in the direction of the Poynting vector, and that reflection of a wave transfers twice the *component* of the momentum of the wave perpendicular to the surface.

The remaining question is, how much momentum does a wave carry, and how can we compute the force exerted by the wave on any given surface? The answer to these two questions – well beyond the scope of this course to *derive* – is that the *momentum density* of an electromagnetic wave is:

$$\vec{g} = \frac{1}{c^2} \vec{S} \quad (893)$$

The magnitude of the momentum Δp transferred to a surface area A that absorbs an electromagnetic wave and that is normal to the wave direction, per unit time, in time Δt , is then all of the momentum in the box of volume $Ac\Delta t$ as usual (we've used this argument many times before) or:

$$\Delta p = \frac{1}{c^2} |\vec{S}| Ac\Delta t \quad (894)$$

If we divide both A and Δt to the left, we get the force per unit area exerted on the surface:

$$P_r = \frac{1}{A} \frac{\Delta p}{\Delta t} = \frac{|\vec{S}|}{c} \quad (895)$$

This is called the **radiation pressure** exerted on the surface by the electromagnetic field, assuming normal incidence and complete absorption of the wave. One then finds the total force the usual way:

$$\vec{F} = A \frac{\vec{S}}{c} \quad (896)$$

in the direction of the wave (the Poynting vector direction itself).

If one 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:

$$\vec{F} = A \frac{\vec{S}}{c} \cos(\theta) \quad (897)$$

but the force is still exerted in the direction of the incident wave.

If the wave is incident on a tipped surface that *reflects* the wave, 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:

$$\vec{F} = 2A \frac{\vec{S}}{c} \cos(\theta) \hat{n} \quad (898)$$

where \hat{n} is a normal unit vector pointing *in* to the surface in question. The momentum density of the incident wave parallel to the surface is unchanged while the momentum density perpendicular to the surface reverses. As noted above, this is an idealization as the reflected wave will always have slightly less energy density than the incident one if the surface itself recoils and gains energy from the wave.

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.

As always, we need to rederive the principle results of the week on our own for homework (has it occurred to you yet that this is one of the things we are doing?). So let's start by using Maxwell's equations to show for a z -directed plane wave (where \vec{E} and \vec{B} are independent of x and y) that:

$$\frac{\partial E_x}{\partial z} = -\frac{\partial B_y}{\partial t} \quad (899)$$

$$\frac{\partial B_y}{\partial z} = -\mu_0 \epsilon_0 \frac{\partial E_x}{\partial t} \quad (900)$$

and

$$\frac{\partial E_y}{\partial z} = \frac{\partial B_x}{\partial t} \quad (901)$$

$$\frac{\partial B_x}{\partial z} = \mu_0 \epsilon_0 \frac{\partial E_y}{\partial t} \quad (902)$$

and from this show that (E_x, B_y) and (E_y, B_x) both satisfy the wave equation for a z -directed wave.

Problem 3.

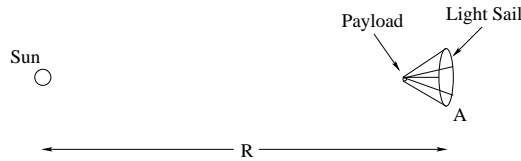
Show that $f(z \pm vt)$ satisfies the wave equation:

$$\frac{\partial^2 f}{\partial x^2} - \frac{1}{v^2} \frac{\partial^2 f}{\partial t^2} \quad (903)$$

Show (by drawing appropriate pictures that convince *you* that it is true so that you *understand* it) that these are left and right propagating waves respectively.

Finally show that $F_0 \cos(kz \pm \omega t)$ is a function that has this form, so that harmonic travelling waves manifestly satisfy the wave equation!

Problem 4.



Some science fiction stories, notably ones by Larry Niven, portray space travel around the solar system occurring with no expenditure of reaction fuel using a *light sail*. A light sail is an enormous, extremely thin, perfectly reflecting mirror arranged like a parachute so that it can "lift" a payload/space capsule attached to the sail by shroud lines. Radiation pressure from sunlight exerts a force on the sail sufficient to lift the mass directly out from the sun, and by altering the angle of the sail one can "tack" in arbitrary directions.

This problem analyzes the plausibility of this proposal. Start by computing the force exerted by sunlight on a perfectly reflecting sail at normal incidence a distance R away from the center of the sun. Note well that a reflecting sail will exert *twice* the force that an absorptive sail would (why?). Next, make a reasonable assumption for the density of the sail material and compute the maximum thickness of a sheet of it that is capable of lifting its own weight against the gravitational pull of the sun. Using this information, *you* decide if the idea of sailing directly away from the sun (with or without a payload) is plausible. Does your answer depend on how far away from the sun you are?

Of course, this simple no-orbit radial model is naive. In reality, the starting and ending point of any journey are *orbits* around the sun; a payload won't fall into the sun even if it has no light sail at all as long as it is in a solar orbit, and one has to do a lot of work on a mass to take it *out* of a solar orbit if it starts in one.

In general, to go from one orbit to another, it suffices to *add energy* (and angular momentum in the proper measure) to the orbiting object (or take them away, of course) in the correct direction using an angled light sail. Making any assumptions that you like, make an argument for or against light sails as a means of moving a significant payload mass

between earth orbit and a lunar orbit, or between earth orbit and an orbit around/near mars without the expenditure of fuel.

In a nutshell, what is the maximum plausible transverse acceleration one can expect to achieve using a light sail of reasonable thickness angled at θ with respect to the sun, for a payload of (say) 1 metric ton (2000 kg)? How large a light sail do you need to achieve that result?

The power output of the sun is 3.8×10^{26} watts, and its mass is 2.0×10^{30} kilograms. If you need it, the mean radius of earth's orbit is $R = 1.5 \times 10^{11}$ meters.

Problem 5.

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 current flows.

- a) Find the net resistance R of the resistor.
- b) Find the current I through the resistor.
- c) Find the electric field inside the resistive material.
- d) Find the magnetic field as a function of distance from the cylinder axis inside the resistive material (assume that its permeability is μ_0).
- e) Evaluate the *Poynting vector* \vec{S} at an arbitrary point on the *cylindrical surface* of the resistor.
- f) Evaluate the *flux of the Poynting vector* through that surface. Simplify it so that is given in terms of I and R . Surprise! The Poynting vector *precisely predicts Joule heating!*

Problem 6.

Let's work out an interesting fact about the solar wind. Consider a spherical grain of dust of radius R with a "reasonable" mass density of 1000 kg per cubic meter (the density of water). Given the mass of the sun (see problem above), your knowledge of G (the gravitational constant) and the insight that the radiation pressure from sunlight is approximately exerted on the transverse cross-sectional area of the sphere πR^2 , determine the radius R_c for which the force exerted by light pressure *away* from the sun *exactly balances* the gravitational force *towards* the sun.

Will particles larger than this (smaller than this) fall into or be pushed away from the sun? Note well that this differential force is exerted no matter how far away from the sun one travels, so particles pushed away are accelerated all the way! This explains why small particles (gas molecules, dust particles) are accelerated *away* from stars, forming a constant "wind" of microparticle radiation.

Problem 7.

Suppose you have a long solenoid (of length L , with $n = N/L$ turns per unit length and radius R) carrying a time varying current $I(t) = I_0(1 - e^{-t/\tau})$.

- Find $B_z(t)$ inside the solenoid.
- Find the induced electrical field at an arbitrary point inside the solenoid (say, at a distance r from its axis).
- Find the magnitude and direction of the Poynting vector on an imagined surface of constant radius just inside the windings at radius R .
- Compute the flux of the Poynting vector into the volume of the solenoid.
- Compute the total magnetic energy of the solenoid, and show that the flux of the Poynting vector equals the rate at which this energy changes.

Problem 8.

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?

Problem 9.

A capacitor consisting of two *circular* conducting disks of radius R is being charged by a steady current I . Find the magnetic and electric fields at an arbitrary point inside the volume of *empty space* between the two plates (using Gauss's Law and Ampere's Law with the Maxwell Displacement Current, respectively). Form the Poynting vector at a point on the "boundary" of the E field, assuming no fringing fields, and integrates the flux of the Poynting vector into the volume of the capacitor. Show that the result equals $P_C = V_C I$, the power being delivered to the capacitor. (Note this problem, the

resistance problem, and the inductance problem are all very similar and have the same purpose – for you to convince yourself that the electromagnetic field carries field energy and is *consistent* with the work-energy theorem implicit in $P = VI$, the rate we do work pushing charge across the potential difference of any device.)

Part I

Optics

Week 11: Light

- The speed of light in a medium is:

$$v_{\text{medium}} = \frac{c}{n} \quad (904)$$

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) $n(\text{red}) < n(\text{violet})$, 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,

$$\theta_i = \theta_\ell \quad (905)$$

- Snell's Law:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \quad (906)$$

- 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:

$$\theta_c = \sin^{-1} \left(\frac{n_1}{n_2} \right) \quad (907)$$

- 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:

$$\begin{aligned}\vec{E}(z, t) &= \frac{\sqrt{2}}{2} E_0 \hat{x} \sin(kz - \omega t \pm \pi/2) + \frac{\sqrt{2}}{2} E_0 \hat{y} \sin(kz - \omega t) \\ \vec{E}(z, t) &= \frac{\sqrt{2}}{2} E_0 (\pm \hat{x} \cos(kz - \omega t) + \hat{y} \sin(kz - \omega t))\end{aligned}\quad (908)$$

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:

$$\vec{E}(z, t) = E_{0x} \hat{x} \sin(kz - \omega t + \delta_x) + E_{0y} \hat{y} \sin(kz - \omega t + \delta_y) \quad (909)$$

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:

$$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \quad (910)$$

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:

$$E_{\text{transmitted}} = \vec{E} \cdot \hat{t} = E_{\text{incident}} \cos(\theta) \quad (911)$$

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:

$$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \quad (912)$$

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:

$$\tan(\theta_b) = \frac{n_2}{n_1} \quad (913)$$

- 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 ($v_s \ll c$):

$$f' = \frac{f}{(1 \mp \frac{v_s}{c})} \quad (914)$$

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 ($v_r \ll c$):

$$f' = f(1 \pm \frac{v_r}{c}) \quad (915)$$

for a receiver moving towards (+) or away from (-) the source.

- Moving Source and Moving Receiver:

Ditto:

$$f' = f \frac{(1 \pm \frac{v_r}{c})}{(1 \mp \frac{v_s}{c})} \quad (916)$$

- 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*.

11.1: 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:

$$v \approx \frac{1}{\sqrt{\epsilon_r\epsilon_0\mu_0}} = \frac{c}{\sqrt{\epsilon_r}} \quad (917)$$

To keep life simple, we take all of the contributing properties of the material and roll them into a single relation:

$$v_{\text{medium}} = \frac{c}{n} \quad (918)$$

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 of 4.3×10^{14} Hz to 7.5×10^{14} Hz, 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*: $n(\omega)$. 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.

11.2: The Law of Reflection

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

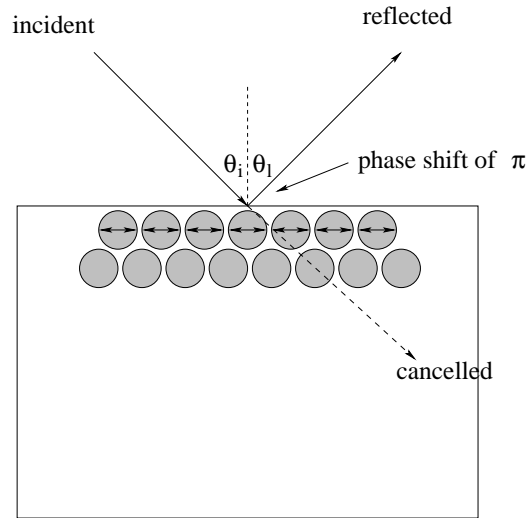


Figure 143: When light is incident on a perfectly reflecting surface, it creates little antennas/sources that radiate the *opposite* field in the direction of the incident field. These antennas cause the light to be reflected at the same angle and with the opposite phase from the surface.

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 perpendicular 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:

$$\theta_i = \theta_r \quad (919)$$

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 by 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.

11.3: Snell’s Law

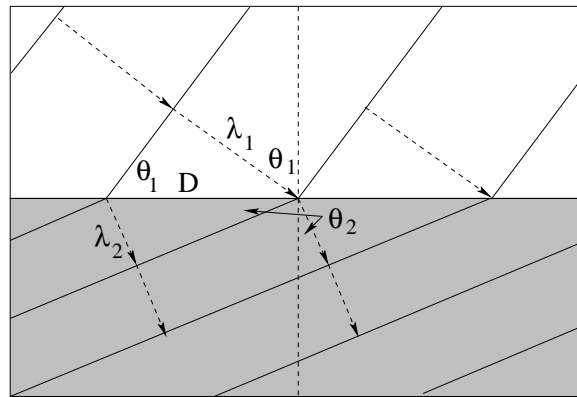


Figure 144: When light is incident on a transparent dielectric surface, it is partially transmitted and partially reflected. Since its *speed* changes, however, the light must *change direction* at the surface as shown.

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 144 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:

$$v = \frac{c}{n} = f\lambda \quad (920)$$

with the same frequency on both sides, it is clear that the wavelength

$$\lambda = \frac{c}{nf} \quad (921)$$

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:

$$D = \frac{\lambda_1}{\sin(\theta_1)} = \frac{\lambda_2}{\sin(\theta_2)} \quad (922)$$

or (substituting from above and cancelling c/f):

$$\frac{1}{n_1 \sin(\theta_1)} = \frac{1}{n_2 \sin(\theta_2)} \quad (923)$$

Inverting, we obtain **Snell's Law**:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \quad (924)$$

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.

Fermat's Principle

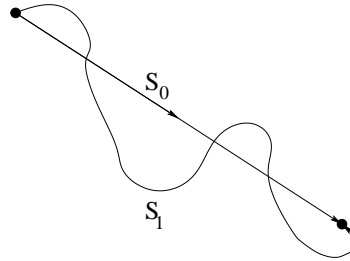


Figure 145: For constant speed, the straight line path between A and B takes the least time.

In figure 145, 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.

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 145 – 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.

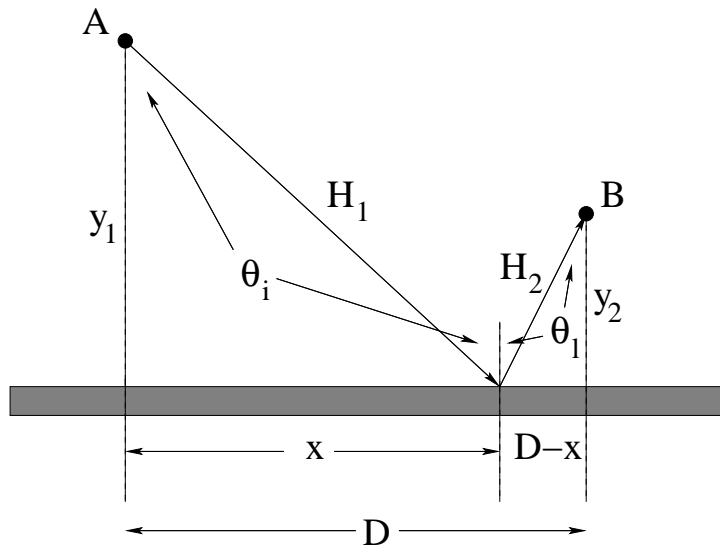


Figure 146: The path with $\theta_i = \theta_r$ is the one with the minimal time when the entire trajectory is otherwise in a single medium with a constant speed.

In figure 146 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:

$$\begin{aligned} H &= H_1 + H_2 \\ &= \{y_1^2 + x^2\}^{\frac{1}{2}} + \{y_2^2 + (D-x)^2\}^{\frac{1}{2}} \end{aligned} \quad (925)$$

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):

$$\frac{dH}{dx} = \frac{\frac{1}{2}2x}{\{y_1^2 + x^2\}^{\frac{1}{2}}} - \frac{\frac{1}{2}2(D-x)}{\{y_2^2 + (D-x)^2\}^{\frac{1}{2}}} = 0 \quad (926)$$

or

$$\sin(\theta_i) = \frac{x}{\sqrt{y_1^2 + x^2}} = \frac{x}{H_1} = \frac{D-x}{H_2} = \frac{(D-x)}{\sqrt{y_2^2 + (D-x)^2}} = \sin(\theta_r) \quad (927)$$

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?

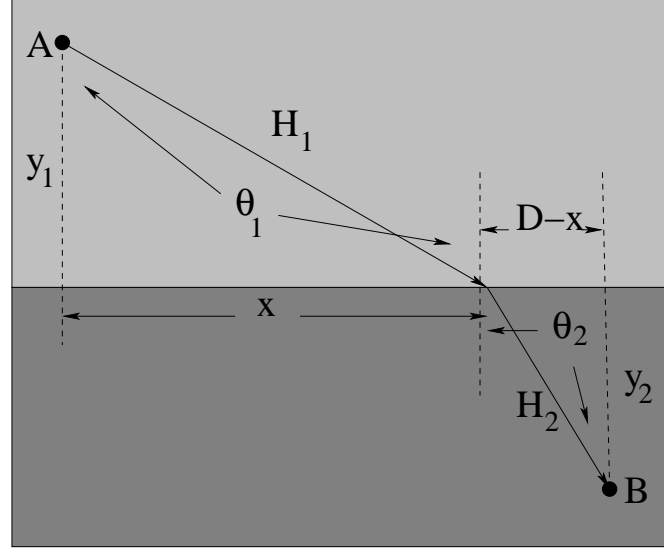


Figure 147: The path with $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$ is the one with the minimal time when the trajectory goes between media n_1 and n_2 where light has distinct speeds. As suggested, one minimizes the time by choosing a trajectory that trades off more distance in the faster medium against less distance in the slower one.

To derive Snell's Law, we need a figure like that one drawn in figure 147. 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 indices for refraction for those media, that is:

$$t_1 = \frac{\sqrt{y_1^2 + x^2}}{v_1} = \frac{n_1 \sqrt{y_1^2 + x^2}}{c} \quad (928)$$

and

$$t_2 = \frac{\sqrt{y_2^2 + (D-x)^2}}{v_2} = \frac{n_2 \sqrt{y_2^2 + (D-x)^2}}{c} \quad (929)$$

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:

$$t = t_1 + t_2 = \frac{n_1 \sqrt{y_1^2 + x^2}}{c} + \frac{n_2 \sqrt{y_2^2 + (D-x)^2}}{c} \quad (930)$$

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:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \quad (931)$$

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!

Total Internal Reflection, Critical Angle

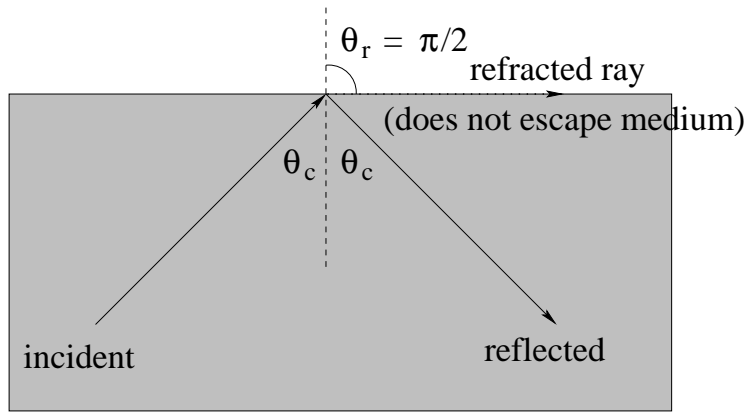


Figure 148: Light travelling from a denser medium to a lighter one is totally internally reflected if $\theta_i \geq \theta_c = \sin^{-1}(\frac{n_1}{n_2})$, corresponding to an angle of refraction of $\pi/2$, where the refracted ray *fails to escape the medium*.

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):

$$n_2 \sin(\theta_c) = n_1 \sin(\pi/2) = n_1 \quad (932)$$

or

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right) \quad (933)$$

If we increase $\theta_i > \theta_c$, 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 (Moissanite) have *extremely large indices of refraction*, roughly $n_d = 2.4$. This makes its critical angle:

$$\theta_{cd} = \sin^{-1} \left(\frac{1}{2.4} \right) = 24.6^\circ \quad (934)$$

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.

Dispersion

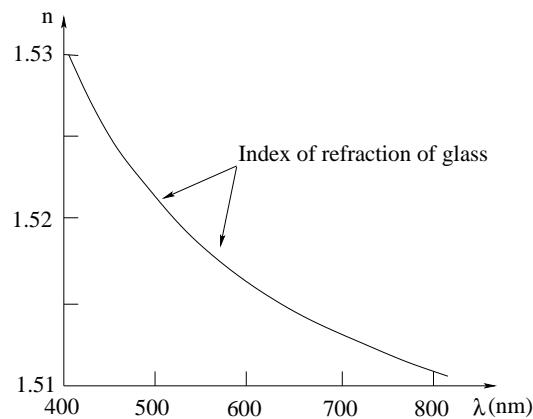


Figure 149: An approximate dispersion curve $n(\lambda)$ for “ordinary” glass. However, distinct glass mixtures can have very different dispersion curves, including ones where n *increases* with increasing wavelength λ (decreases with frequency).

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 149. 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**.

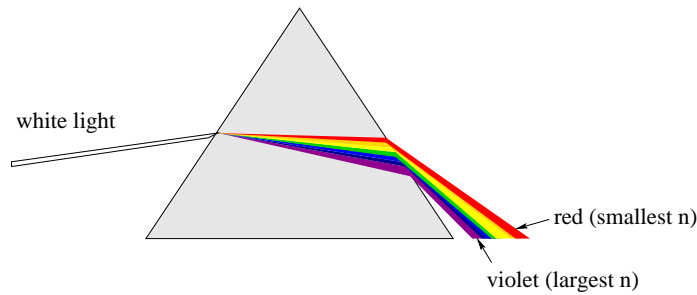


Figure 150: A prism causes violet light to be bent more than red light at each interface, splitting up the originally white incident light into a full spectrum.

In figure 150 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*¹⁰⁶. 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.

11.4: 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). We describe the behavior of the two components of the *electric* field component for a given fixed harmonic frequency as the *polarization* of the harmonic wave. There are several ways to describe the polarization, and several physical processes produce polarized light.

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.

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

¹⁰⁶Or, 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...

examples of linearly polarized light propagating in the z -direction with frequency ω :

Light linearly polarized in the x -direction:

$$\vec{E}(z, t) = E_{0x} \hat{x} \sin(kz - \omega t) \quad (935)$$

(The associated magnetic field *must* be:

$$\vec{B}(z, t) = B_{0y} \hat{y} \sin(kz - \omega t) = \frac{E_{0x}}{c} \hat{y} \sin(kz - \omega t) \quad (936)$$

according to the rules derived in the previous chapter, because

$$|\vec{B}| = \frac{|\vec{E}|}{c} \quad (937)$$

and because

$$\hat{x} \times \hat{y} = \hat{z} \quad (938)$$

in the Poynting vector.)

Light linearly polarized in the y -direction:

$$\vec{E}(z, t) = E_{0y} \hat{y} \sin(kz - \omega t) \quad (939)$$

(The associated magnetic field *must* be:

$$\vec{B}(z, t) = -B_{0x} \hat{x} \sin(kz - \omega t) = -\frac{E_{0y}}{c} \hat{x} \sin(kz - \omega t) \quad (940)$$

according to the rules derived in the previous chapter, because

$$\hat{y} \times -\hat{x} = \hat{z} \quad (941)$$

in the Poynting vector.)

Finally, light linearly polarized along the line at $\pi/4$ above the x -axis is::

$$\vec{E}(z, t) = \frac{\sqrt{2}}{2} E_0 \hat{x} \sin(kz - \omega t) + \frac{\sqrt{2}}{2} E_0 \hat{y} \sin(kz - \omega t) \quad (942)$$

The amplitude of the electric field is E_0 (why?). What must the direction and magnitude of the associated magnetic field?

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:

$$\begin{aligned} \vec{E}(z, t) &= \frac{\sqrt{2}}{2} E_0 \hat{x} \sin(kz - \omega t \pm \pi/2) + \frac{\sqrt{2}}{2} E_0 \hat{y} \sin(kz - \omega t) \\ \vec{E}(z, t) &= \frac{\sqrt{2}}{2} E_0 (\pm \hat{x} \cos(kz - \omega t) + \hat{y} \sin(kz - \omega t)) \end{aligned} \quad (943)$$

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).

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:

$$\vec{E}(z, t) = E_{0x} \hat{x} \sin(kz - \omega t + \delta_x) + E_{0y} \hat{y} \sin(kz - \omega t + \delta_y) \quad (944)$$

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 with 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>

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):

$$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \quad (945)$$

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:

$$E_{\text{transmitted}} = \vec{E} \cdot \hat{t} = E_{\text{incident}} \cos(\theta) \quad (946)$$

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:

$$I = \left| \frac{1}{2\mu_0} \vec{E} \times \vec{B} \right| = \frac{1}{2\mu_0 c} E^2 \quad (947)$$

The intensity is directly proportional to the electric field amplitude, squared, so that:

$$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \quad (948)$$

This result is known as **Malus's law**.

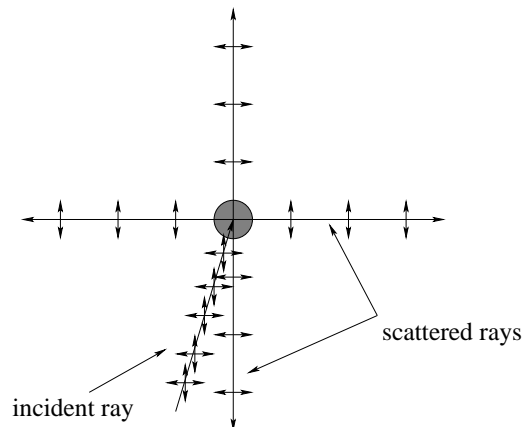


Figure 151: The scattering of initially unpolarized light by a molecule or dust particle. Note that the polarization is perpendicular to the *plane of scattering* for each of the possible outgoing directions.

Polarization by Scattering

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!*

Polarization by Reflection

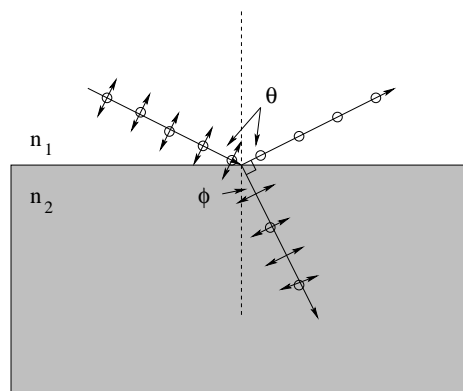


Figure 152: The scattering of initially unpolarized light by reflection off of a plane surface between two dielectric media at the *Brewster angle* that produces complete polarization of the reflected ray. Note that the polarization of all reflected rays incident on the surface at an angle is *parallel 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:

$$n_1 \sin(\theta) = n_2 \sin(\phi) \quad (949)$$

is Snell's law, but clearly:

$$\phi = \frac{\pi}{2} - \theta \quad (950)$$

so that:

$$\sin(\phi) = \sin(\pi/2 - \theta) = \cos(\theta) \quad (951)$$

and *Brewster's formula*:

$$\tan(\theta_b) = \frac{n_2}{n_1} \quad (952)$$

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.

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.

11.5: 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***¹⁰⁷.

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

¹⁰⁷At 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

observes. If both motions are occurring, both shifts occur as a product. We show the picture and quick derivation of each possibility below.

Moving Source

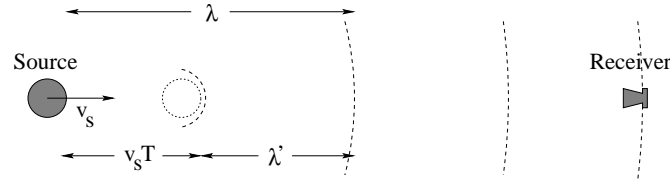


Figure 153: Wave geometry for Doppler shift of moving source.

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 $v_s T$, reducing the distance at the time of the next front to $\lambda' = \lambda - v_s T$. 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:

$$\begin{aligned}
 \lambda' &= \lambda - v_s T \\
 cT' &= cT - v_s T \\
 T' &= T \left(1 - \frac{v_s}{c}\right) \\
 \frac{1}{T'} &= \frac{1}{T} \frac{1}{\left(1 - \frac{v_s}{c}\right)} \\
 f' &= \frac{f}{\left(1 - \frac{v_s}{c}\right)}
 \end{aligned} \tag{953}$$

For a source moving *away* from the receiver the algebra and picture is the same, but the wavelength $\lambda' = \lambda + v_s T$ is *increased*, so that:

$$f' = \frac{f}{\left(1 \mp \frac{v_s}{c}\right)} \tag{954}$$

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.

Moving Receiver

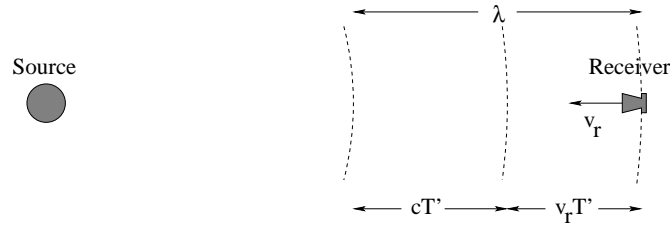


Figure 154: Wave geometry for Doppler shift of a moving receiver.

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 $\lambda = cT = (c + v_r)T'$:

$$\begin{aligned}
 cT &= (c + v_r)T' \\
 T &= \left(1 + \frac{v_r}{c}\right)T' \\
 \frac{1}{T'} &= \frac{1}{T}\left(1 + \frac{v_r}{c}\right) \\
 f' &= f\left(1 + \frac{v_r}{c}\right)
 \end{aligned} \tag{955}$$

As before, if the receiver is moving away, it decreases f' instead of increasing it, so that the general rule is:

$$f' = f\left(1 \pm \frac{v_r}{c}\right) \tag{956}$$

for a receiver moving towards (+) or away from (-) the source.

Moving Source and Moving Receiver

The rule is just the product of the two rules:

$$f' = f \frac{\left(1 \pm \frac{v_r}{c}\right)}{\left(1 \mp \frac{v_s}{c}\right)} \tag{957}$$

It is interesting to note that if a source is moving at the speed of light (where these expressions are no longer valid, alas, although they still capture *part* of the shift) the frequency f' goes to *infinity*. This divergence occurs in the relativistic expression as well, and is the moral equivalent of a *sonic boom* only with light.

Although particles cannot go faster than light *in a vacuum*, this is actually a physical possibility *inside a medium*. 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 particle 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.

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 Snell's Law. You may use any method you like (there are several) but the way it was done in class is probably the easiest).

Problem 3.

Derive the 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 in a vacuum, where the upper signs in both case refer to approach and the lower signs recession. Note well that this is how the radar guns police use to trap speeder work, how "doppler radar" used by weather forecasters works that measure the wind speed of storms and can detect the occurrence of tornados, and is a technology used in a variety of medical imaging techniques including e.g. ultrasound.

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 pictures representing:

- Polarization by scattering
- Polarization by absorption
- Polarization by reflection

These are a mnemonic device for the formulas and help you understand why the transmission axis of polarizing sunglasses is *vertical* (to block reflected glare and scattered skylight, both predominantly polarized parallel to the ground).

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).

Problem 8.

Suppose a layer of oil $n_o = 5/4$ is floating on water $n_w = 4/3$, that in turn is on a piece of glass $n_g = 3/2$. Show that the critical angle for the glass is not changed by the *combined system* of layers of water and oil; that rays incident on the glass-water interface at or above the critical angle for glass-air alone do not escape the final layer of oil.

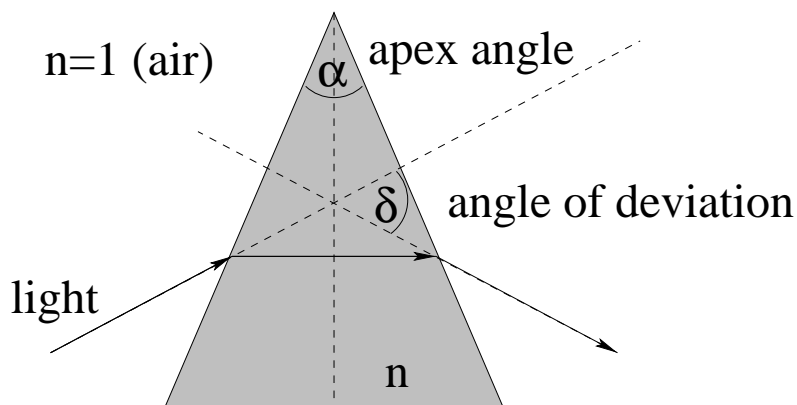
Problem 9.

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.

Problems 10 and 11 on next page

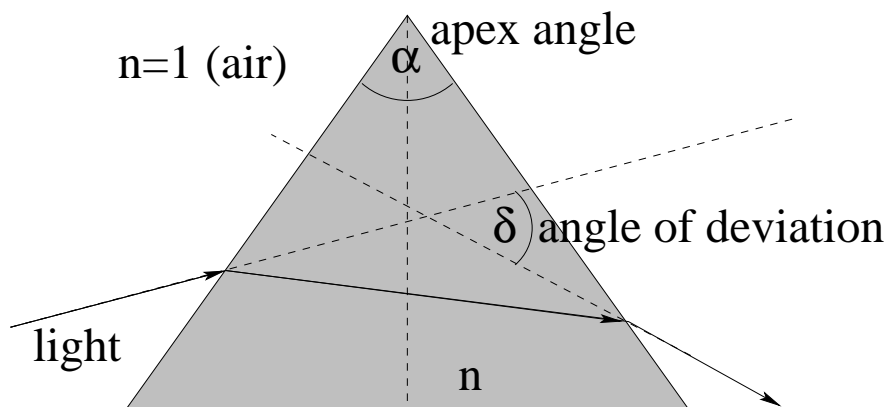
Problem 10.



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.

Week 12: 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:

$$d = \frac{1}{f} \quad (958)$$

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:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad (959)$$

- The transverse magnification of a simple mirror (or lens) is defined by the ratio of the image height y' to the object height y :

$$m = \frac{y'}{y} = -\frac{s'}{s} \quad (960)$$

- 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:

$$f = \frac{r}{2} \quad (961)$$

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**:

$$\frac{1}{f} = (n_2 - n_1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (962)$$

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:

$$M = \frac{x_{np}}{f} \quad (963)$$

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:

$$M = -\frac{f_o}{f_e} \quad (964)$$

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 undesirable. 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:

$$M_o = -\frac{\ell}{f_o} \quad (965)$$

The magnification of the eyepiece (simple magnifier) is:

$$M_e = \frac{x_{np}}{f_e} \quad (966)$$

The overall magnification is therefore:

$$M_{tot} = -\frac{\ell x_{np}}{f_o f_e} \quad (967)$$

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 undesirable. A “Galilean” microscope does not invert the image.

12.1: Vision and Plane Mirrors

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.

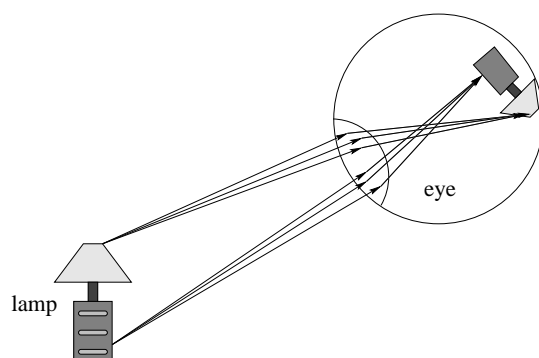


Figure 155: How the eye sees an object. Light diverging from points on the surface of the object are focused onto the retina of the eye, where they form an *image* of the object that the retina converts into neural impulses and your brain converts into perception.

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.

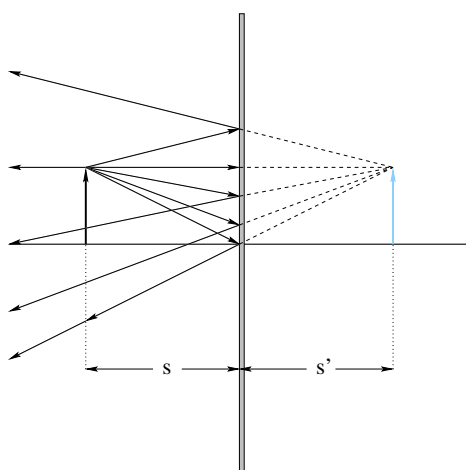


Figure 156: The geometry of forming an image in a plane mirror.

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.

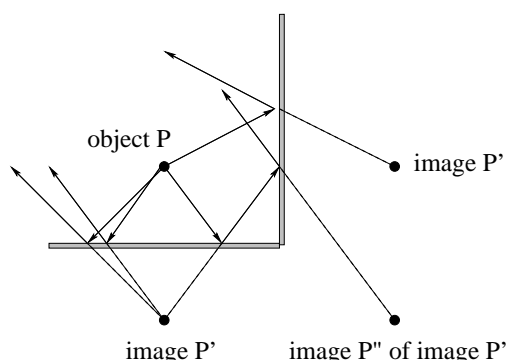


Figure 157: Two mirrors create an image of an image. Only a few of the many rays are drawn – copy the picture and fill in more yourself.

12.2: 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 (156).

In figure (158) we see a *concave spherical mirror*, which we will also call a *converging* mirror or a *positive* mirror¹⁰⁸. 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

¹⁰⁸For those who have concave/convex dyslexia, remember that *concave* is like a cave, and curves inward, while *convex* is nothing at all like a vex. What is a vex, anyway?

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 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.

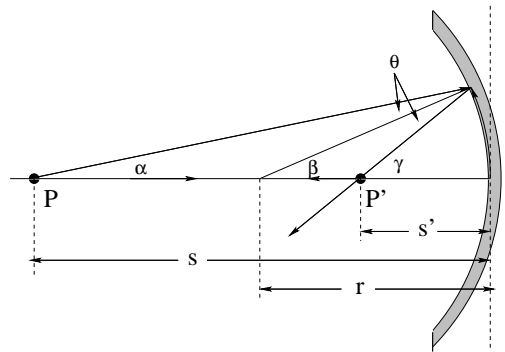


Figure 158: The geometry of forming an image in a concave mirror.

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:

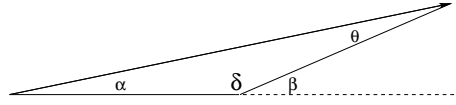
$$\alpha \approx \frac{l}{s} \quad (968)$$

$$\beta = \frac{l}{s} \quad (969)$$

$$\gamma \approx \frac{l}{s'} \quad (970)$$

(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 (159)). We can

Figure 159: $\alpha + \theta = \beta$.

easily see that $\alpha + \theta + \delta = \pi$. But we can *also* see that $\delta + \beta = \pi$. Therefore:

$$\alpha + \theta = \beta \quad (971)$$

and similarly (considering the other triangle involving β and θ)

$$\beta + \theta = \gamma \quad (972)$$

If we eliminate θ , we get:

$$\alpha + \gamma = 2\beta \quad (973)$$

Finally, if we substitute in all of the small angle approximations and cancel l , we get:

$$\frac{1}{s} + \frac{1}{s} = \frac{2}{r} \quad (974)$$

As we move the object back farther and farther from the mirror (let $s \rightarrow \infty$) 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:

$$f = \frac{r}{2} \quad (975)$$

and

$$\frac{1}{s} + \frac{1}{s} = \frac{1}{f} \quad (976)$$

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 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:

$$d = \frac{1}{f} \quad (977)$$

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.

It is possible to use the same inverse length units to write the thin lens/mirror equation above. If we define $x = 1/s$, $x' = 1/s'$, then:

$$x + x' = d \quad (978)$$

is the *direct* (instead of reciprocal) rule. Note well that the ranges of x , x' , and d have a very different meaning. $d = 0$ means a focal length of $\pm\infty$, a flat mirror (or non-focusing lens). $x = 0$ is similarly $s = \pm\infty$, generally $+\infty$. Here it is quite easy to see how and when x and x' change sign if either one of them is larger than d .

However, this is not necessarily easier to use for the purposes of computation, as one still (ultimately) has to do the same algebra to actually compute s and/or s' .

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.

12.3: 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.

In this figure, $f = 10$ cm, $s = 25$ cm. Therefore:

$$\begin{aligned}\frac{1}{25} + \frac{1}{s'} &= \frac{1}{10} \\ \frac{1}{s'} &= \frac{1}{10} - \frac{1}{25} \\ \frac{1}{s'} &= \frac{1.5}{25} \\ s' &= \frac{25}{1.5} = 16.7 \text{ cm}\end{aligned}\tag{979}$$

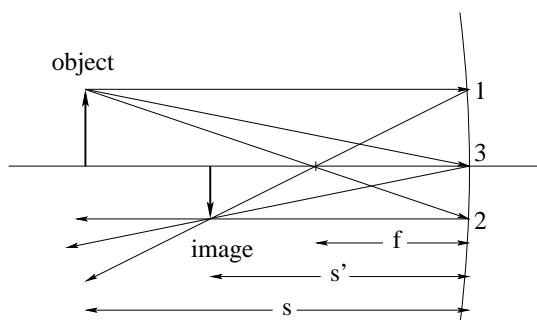
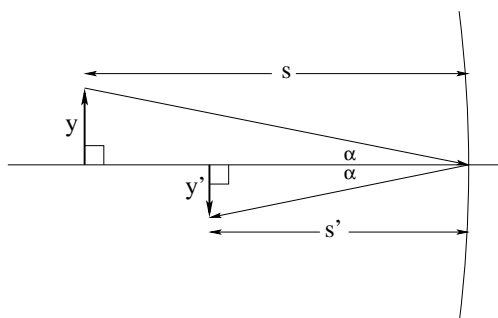
Figure 160: Converging mirror with $s = 25 > f = 10$.

Figure 161: Transverse magnification can be determined from the two right triangles formed with the central ray as a hypotenuse.

To compute the magnification of the image formed above, we note that:

$$\tan(\alpha) = -\frac{y}{s} = \frac{y'}{s'} \quad (980)$$

(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:

$$m = \frac{y'}{y} = -\frac{s'}{s} \quad (981)$$

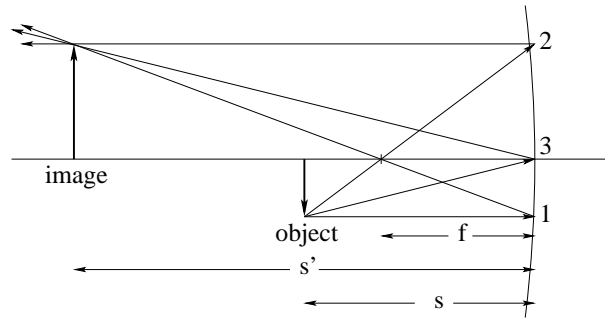
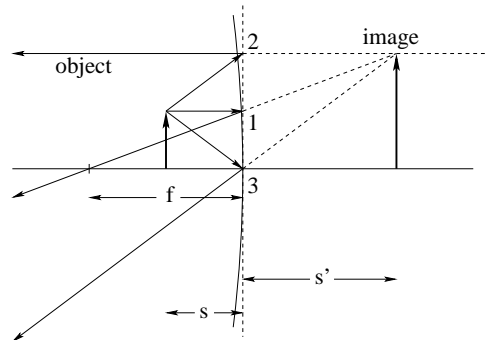
This expression is valid for all images obtained for any ideal lens or mirror.

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 (162), 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.

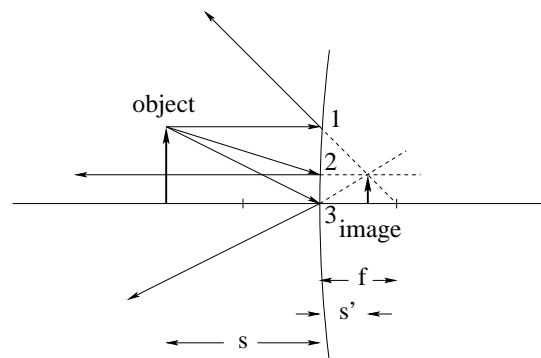
As before, $1/s' = 1/10 - 1/15 = 1/30$ so $s' = 30$ cm. The magnification is $m = -s'/s = -30/10 = -3$. The image is again real and inverted (relative to the object), but in this case the image is larger than the object.

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 (162). 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.

Figure 162: Converging mirror with $2f = 20 > s = 15 > f = 10$.Figure 163: Converging mirror with $s = 5 < f = 10$.

In figure (163) 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 164: 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.

12.4: Lenses

A spherical lensing surface between two different media with different indices of refraction are drawn in figure (165).

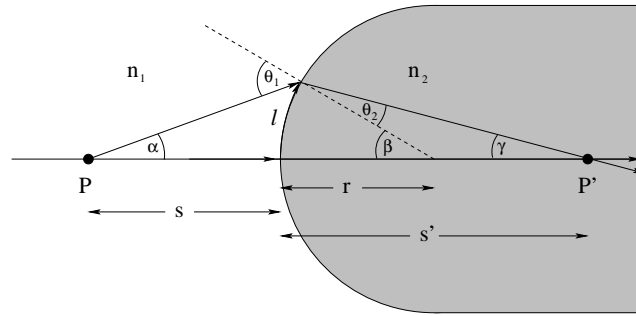


Figure 165: Diagram that shows how a spherical lens creates an image via refraction.

As was the case for the mirror, the three angles α , β , and γ in the small angle approximation can be written as:

$$\alpha \approx \frac{\ell}{s} \quad (982)$$

$$\beta = \frac{\ell}{r} \quad (983)$$

$$\gamma \approx \frac{\ell}{s'} \quad (984)$$

We also have *Snell's law* for the (small) angles θ_1 and θ_2 :

$$n_1 \theta_1 \approx n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \approx n_2 \theta_2 \quad (985)$$

so

$$\theta_2 = \frac{n_1}{n_2} \theta_1. \quad (986)$$

Using triangle rules like the ones above, we also get:

$$\theta_1 = \alpha + \beta \quad (987)$$

and

$$\beta = \theta_2 + \gamma \quad (988)$$

Eliminating θ_2 , this becomes:

$$\beta = \frac{n_1}{n_2} \theta_1 + \gamma \quad (989)$$

If we multiply both sides by n_2 and substitute θ_1 from the first equation, this becomes:

$$n_2 \beta = n_1 \alpha + n_1 \beta + n_2 \gamma \quad (990)$$

or

$$n_1\alpha + n_2\gamma = (n_2 - n_1)\beta \quad (991)$$

We substitute in the small angle formulas and cancel l to get:

$$\frac{n_1}{s} + \frac{n_2}{s'} = (n_2 - n_1)\frac{1}{r} \quad (992)$$

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:

$$\frac{1}{s} + \frac{n}{s'} = (n - 1)\frac{1}{r} \quad (993)$$

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 (166). The first surface struck by light from an object (presumed coming in from the left) has

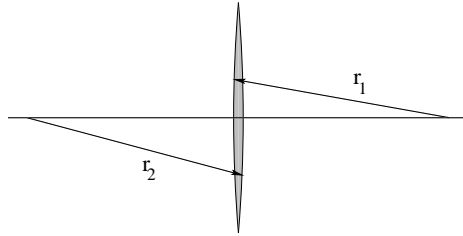


Figure 166: Geometry of a thin lens surrounded by air.

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:

$$\frac{1}{s} + \frac{n}{s'} = (n - 1)\frac{1}{r_1} \quad (994)$$

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:

$$\frac{-n}{s'} + \frac{1}{s''} = (1 - n)\frac{1}{r_2} \quad (995)$$

If we add these two formulae, the s' term cancels and, we get:

$$\frac{1}{s} + \frac{1}{s''} = (n - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = \frac{1}{f} \quad (996)$$

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**:

$$\frac{1}{f} = (n - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \quad (997)$$

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.

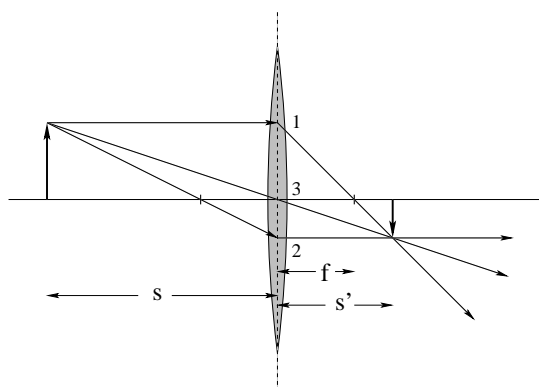


Figure 167: A converging lens with focal length of 10 cm and an object at $s = 30$ cm.

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 (168), the image

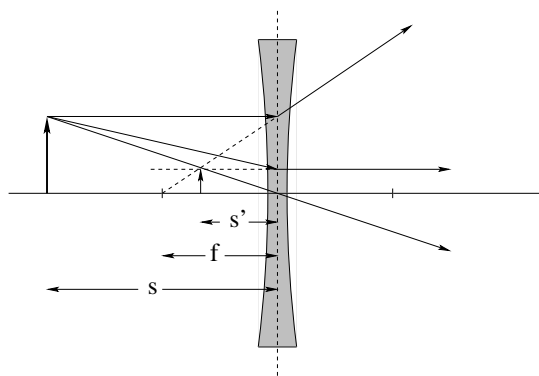


Figure 168: A diverging lens with focal length of -10 cm and an object at $s = 20$ cm.

is virtual, erect, and smaller than the original object. Show (from the numbers and thin lens formula) that $s' = -6.67$ cm and that $m = 1/3$.

12.5: The Eye

The eye is roughly spherical and approximately one inch in diameter. Figure (169) 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...

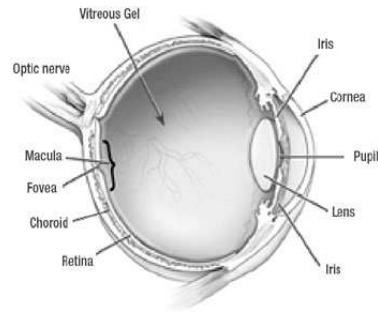


Figure 169: A simplified anatomical diagram of the human eye.

- **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.

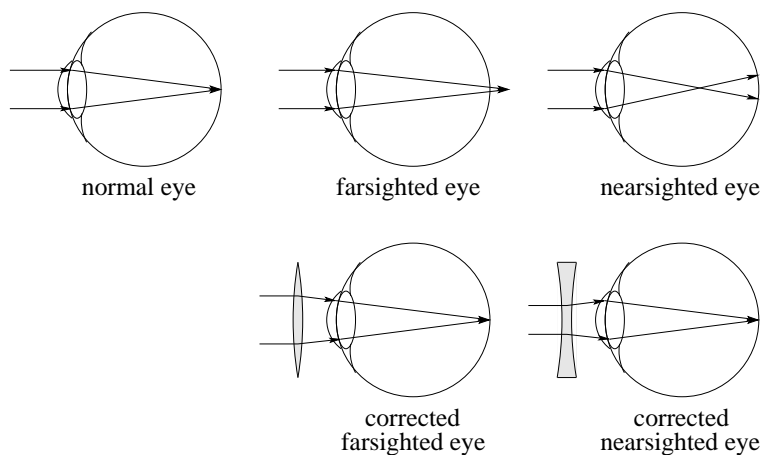


Figure 170: The focal length of the relaxed (combined) lensing acting of the eye for a normal eye, a farsighted eye (hyperopia), and a nearsighted eye (myopia).

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 $\frac{1}{0.025} = 40.00\text{d}$. When viewing less distant objects, accommodation *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 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 not 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.

12.6: Optical Instruments

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.

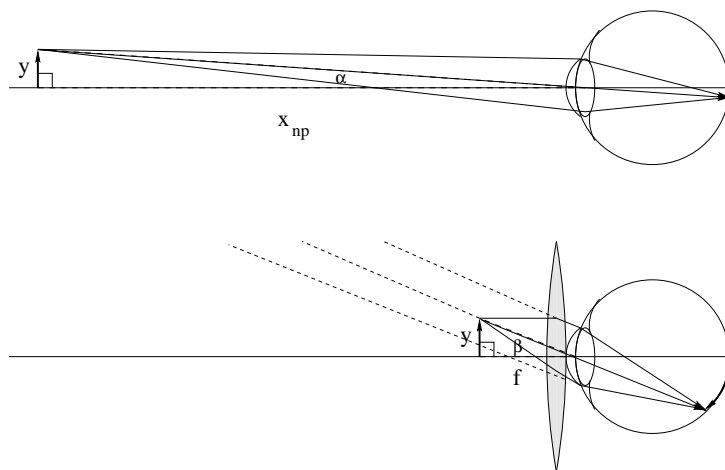


Figure 171: A converging lens used as a simple magnifier.

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 (171) 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 accommodated eye, it subtends an angle of α , where:

$$\alpha \approx \tan(\alpha) = \frac{y}{x_{np}} \quad (998)$$

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:

$$\beta \approx \tan(\beta) = \frac{y}{f} \quad (999)$$

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:

$$M = \frac{\beta}{\alpha} = \frac{x_{np}}{f} \quad (1000)$$

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.

Telescope

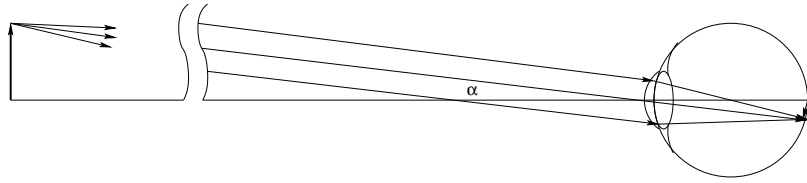


Figure 172: An regular (inverting) telescope.

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 (172) 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 .

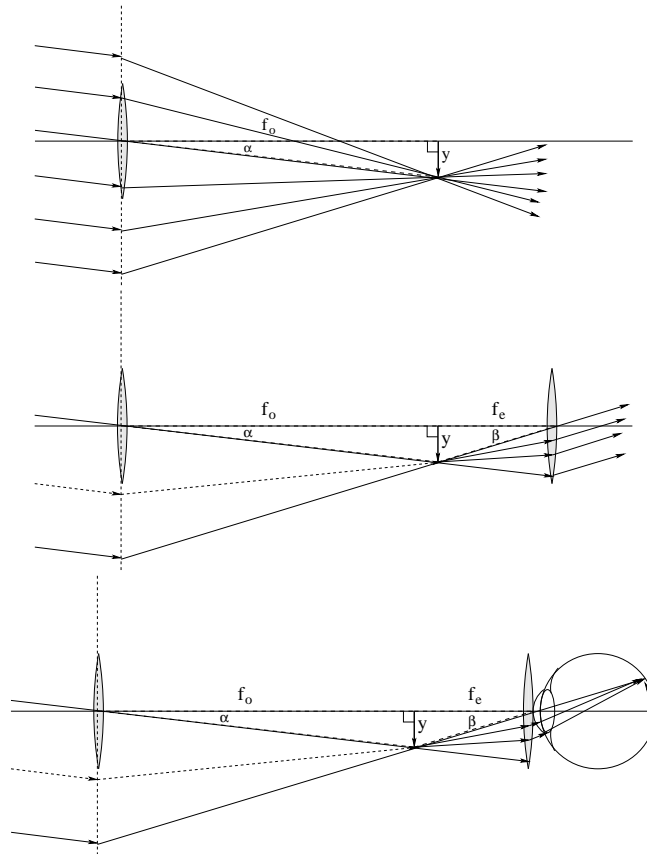


Figure 173: An regular (inverting) telescope.

To magnify our view of this object, we begin by inserting 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 (173) 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.

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 is much steeper. We can easily compute the angular magnification of this telescope by noting that:

$$\alpha \approx \tan(\alpha) = -\frac{y}{f_o} \quad (1001)$$

and

$$\beta \approx \tan(\beta) = \frac{y}{f_e} \quad (1002)$$

so that

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \quad (1003)$$

In the final panel, 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** lens 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*.

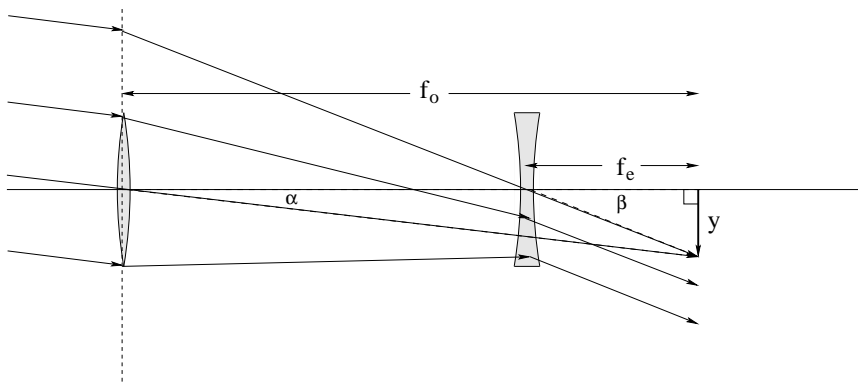


Figure 174: A “Galilean” telescope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the distant objects *erect* instead of inverted.

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 (174), the angular

magnification of a Galilean telescope is still:

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \quad (1004)$$

(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 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.

Microscope

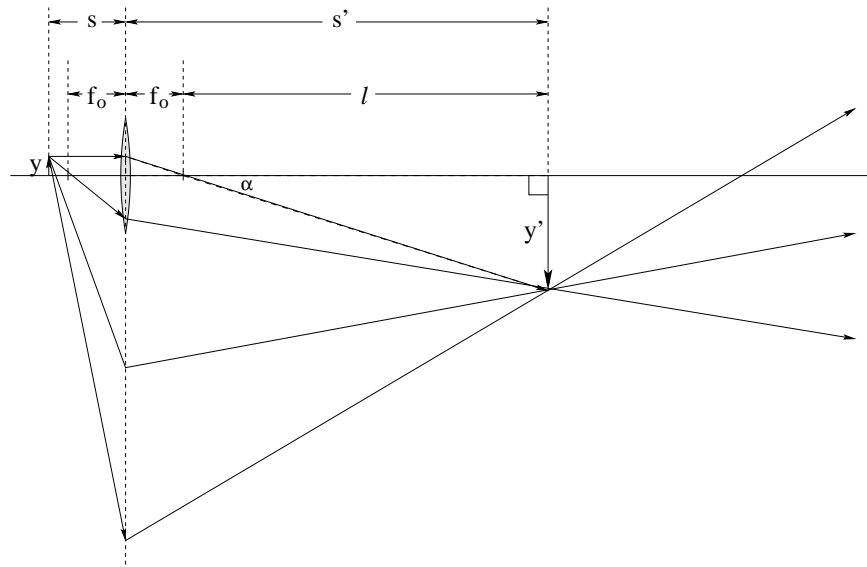


Figure 175: The first magnification stage of a compound microscope brings a *small* object just outside of the focal point of the objective lens into focus as a *real, magnified image* at the end of the **tube length** l . By comparing the two dashed similar triangles, one can see that the first stage 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 (175).

The magnification of the objective stage is:

$$M_o = -\frac{\ell}{f_o} = -\frac{f_o + l}{s} \quad (1005)$$

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.

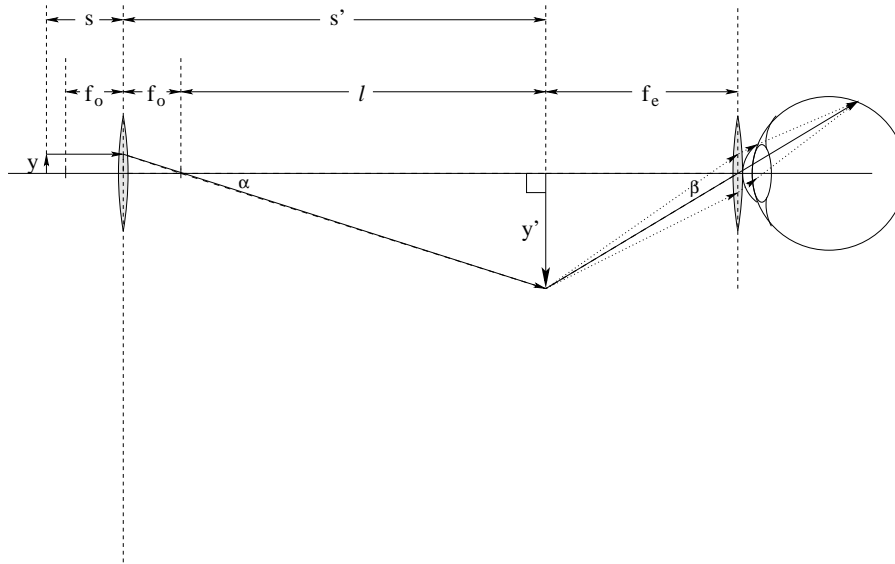


Figure 176: The second magnification stage of a compound microscope brings the *highly magnified* image from the objective stage close to the eye by functioning as a simple magnifier. By bringing the virtual image in from x_{np} to f_e it magnifies it by an additional factor of $\frac{x_{np}}{f_e}$.

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:

$$M_e = \frac{x_{np}}{f_e} \quad (1006)$$

which yields an overall magnification for the two stages working together of:

$$M_{tot} = -\frac{\ell}{f_o f_e} \frac{x_{np}}{f_e} \quad (1007)$$

As we noted and can see in figure (177) 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 undesirable) 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*

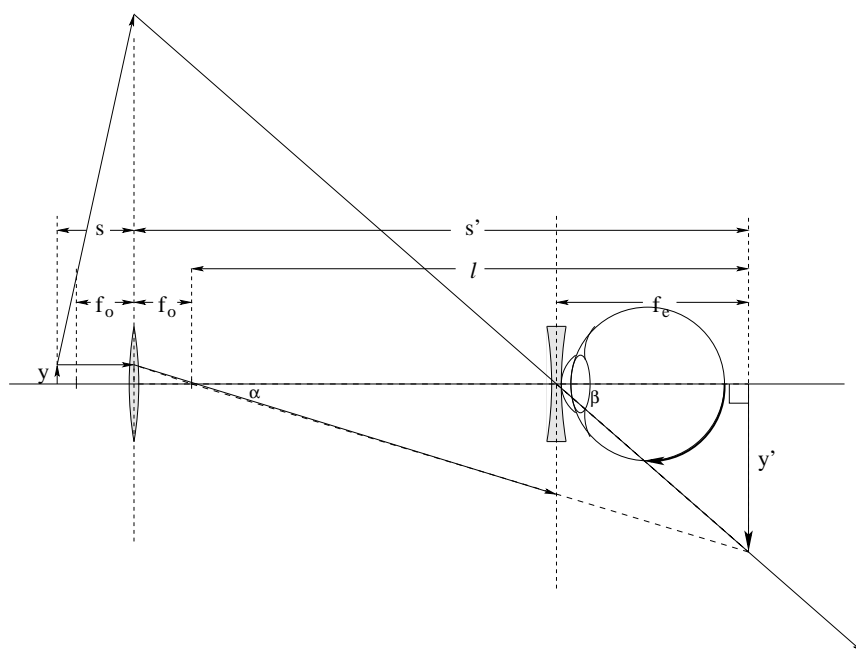


Figure 177: A “Galilean” microscope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the tiny objects *erect* instead of inverted. As always, we use a “central” ray for the second lens that is deflected at the plane of the first lens *as if* it passes through both lenses to find the location and size of the final image.

adequate to observe e.g. blood cells, bacteria, the cellular structure of plant and 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 Enlightenment, a time of intense discovery and invention that ultimately ushered in the rational modern world of today.

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.

Derive the 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 locate its point-image, then doing some work with triangles and the small angle approximation.

Problem 3.

Produce ray diagrams for both lenses and mirrors for all permutations of the following data: $f = 10$ cm. $f = -10$ cm. $s = 10, 20, 40, 60$ cm. In all cases locate the image (give s'), find the magnification m , and indicate whether the image is erect or virtual.

Problem 4.

Prove that the *lateral magnification* of an object is:

$$m_l = \frac{\Delta s'}{\Delta s} = \frac{s'^2}{s^2} \quad (1008)$$

I'd "suggest" that you think about your friend, the *binomial expansion*, when solving this problem. Is the image "inverted"?

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 fish's eye has a focal length of 1 cm in water (which is just the distance from the lens to the fish's retina, of course). Is its focal length in air longer or shorter? Don't just answer with a guess – you need to make a complete argument based on the lens-maker's formula or Snell's law directly, supported by pictures. Is the fish nearsighted or farsighted in air? Conversely, if you open your eyes underwater (and have normal vision in air) are you nearsighted or farsighted?

Problem 7.

Draw ray diagrams and derive the magnification for: The standard telescope and the Galilean telescope (one with an eyepiece lens with a negative focal length). Show that the latter permits one to view the final image at infinity erect instead of inverted.

Problem 8.

Draw ray diagrams and derive the magnification for: The standard microscope (with tube length ℓ) and the "Galilean" microscope (one with an eyepiece lens with a negative focal length). Show that the latter permits one to view the final image at infinity erect instead of inverted.

Problem 9.

- a) Draw a ray diagram for the simple magnifier, deriving its (angular) magnification in the standard picture.
- b) Solve for where one has to locate the object to form a virtual, erect image at the **near point of the eye** x_{np} as viewed through the magnifier.
- c) What is the overall (angular) magnification of the image now (with the image located at x_{np})?

Problem 10.

From the previous problem, you saw that if one places the object viewed with a simple magnifier at a position that isn't exactly at focal point of the lens, one can achieve a slightly greater angular magnification (at the expense of having to use accommodation in order to view the final image at the near point of the eye instead of at infinity). Both the microscope and telescope above use the eyepiece lens as a simple magnifier to view a real image.

Based on your result, by roughly what ***fraction*** do you think you can increase their effective magnification if you locate the final image at the near point of the eye? Note

that you can solve this problem by redoing the diagrams and computation of the overall magnification, or by using your result from the previous problem to estimate the fractional increase in magnification in terms of x_{np} and f_e . Both will help you understand everything better.

Week 13: 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 they 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*¹⁰⁹ 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 τ_{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.

- 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

¹⁰⁹Wikipedia: [http://www.wikipedia.org/wiki/Coherence_\(physics\)](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.

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:

$$d \sin(\theta) = m\lambda \quad \text{Maxima}$$

$$d \sin(\theta) = (m + \frac{1}{2})\lambda \quad \text{Minima}$$

with $m = 0, \pm 1, \pm 2, \pm 3, \dots$

- **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 $m = 0, 1, 2, \dots$. 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!**) $n = \cancel{0}, 1, 2, \dots, N-1, \cancel{N}, N+1, N+2, \dots, 2N-1, \cancel{2N}, 2N+1, \dots$. 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!**) $n = \cancel{0}, 1, 2, \dots, N-1, \cancel{N}, N+1, N+2, \dots, 2N-1, \cancel{2N}, 2N+1, \dots$. 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 $m = 0, 1, 2, \dots$

- 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 λ , 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!**) $m = 1, 2, 3, \dots$. 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 interference 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 independently.

- **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 interested in using this when the angles involved are very small, in which case we can write:

$$\theta_{\min} = 1.22 \frac{\lambda}{D}$$

- 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:

$$\alpha_{\text{incidence}} > \theta_{\min} = 1.22 \frac{\lambda}{D}$$

- **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 reflects 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 n_2** . The heuristic rule is then:

$$2d = m\lambda' = m\frac{\lambda}{n_2} \quad \text{Maxima}$$

$$2d = (m + \frac{1}{2})\lambda' = (m + \frac{1}{2})\frac{\lambda}{n_2} \quad \text{Minima}$$

with $m = 0, \pm 1, \pm 2, \pm 3, \dots$ 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:**

$$2d = (m + \frac{1}{2})\lambda' = (m + \frac{1}{2})\frac{\lambda}{n_2} \quad \text{Maxima}$$

$$2d = m\lambda' = m\frac{\lambda}{n_2} \quad \text{Minima}$$

with $m = 0, \pm 1, \pm 2, \pm 3, \dots$

- 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 nonreflecting coatings for glass and lenses to maximize their light transmission.

13.1: Harmonic Waves and Superposition

Several weeks ago we learned about **harmonic waves**, solutions to the wave equation of the general form (in one dimension):

$$\vec{E}(x, t) = E_0 \hat{e} \sin(kx - \omega t) \quad (1009)$$

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:

$$\vec{E}(r, t) = E_0 \frac{a}{r} \hat{e} \sin(kr - \omega t) \quad (1010)$$

(where $|\vec{E}(a, t)| = E_0$ 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:

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \quad (1011)$$

points in the direction of propagation. Finally, don't forget that the (time averaged) intensity of the wave is:

$$I_0 = \langle |\vec{S}| \rangle_{\text{av}} = \frac{1}{2\mu_0} E_0 B_0 = \frac{1}{2\mu_0 c} E_0^2 \quad (1012)$$

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:

$$\lambda' = \frac{\lambda}{n} \quad (1013)$$

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 atoms that are giving off the light are very weakly correlated with one another.

13.1.1: Hot Sources and Wave Coherence

Although we've seen that Maxwell's equations in free space become the electromagnetic wave equation (so that light is plausibly an 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 to 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):

$$\vec{E}(z, t) = E_{0x} \hat{x} \sin(kz - \omega t) \quad (1014)$$

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* coherent¹¹⁰:

$$\vec{E}(z, t) = E_{0x} \hat{x} \sin(kz - \omega(t)t + \phi(t)) \quad (1015)$$

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.

$$\omega_{\text{avg}} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \omega(t) dt \quad (1016)$$

¹¹⁰And 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 ω_{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.

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 range¹¹¹, 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.

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”¹¹² 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***¹¹³! 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*.

¹¹¹We speak of “line broadening” and the “natural width” of spectral lines to acknowledge or quantify this.

¹¹²In quotes because the fourier transform of a harmonic wave with random phase interruption is no longer sharp or monochromatic.

¹¹³All of this is proven in more advanced mathematical treatments.

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*: τ_{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:

$$\tau_{\text{coh}} \approx 10^{-12} \text{ seconds} \quad (1017)$$

(given frequencies in the range of 10^{14} to 10^{15} cycles per second).

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:

$$L_{\text{coh}} = c\tau_{\text{coh}} \approx 10^{-4} \text{ meters} \quad (1018)$$

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 frequencies 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 holography¹¹⁴, the ring laser gyroscope¹¹⁵ and laser interferometry¹¹⁶.

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

¹¹⁴Wikipedia: <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.

¹¹⁵Wikipedia: [http://www.wikipedia.org/wiki/ring laser gyroscope](http://www.wikipedia.org/wiki/ring%20laser%20gyroscope).

¹¹⁶Wikipedia: <http://www.wikipedia.org/wiki/interferometry>.

question, and even white light from hot sources – as a mixture of many frequencies that are *all* coherent over similar L_{coh} – can be locally sufficiently coherent to support e.g. thin film interference in all of the colors/frequencies independently.

13.1.2: 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_{coh} of the light, we will then have to repeatedly evaluate below sums such as (for a single polarization component of the wave):

$$E_{\text{tot}} = E_1 \sin(kx - \omega t + \delta_1) + E_2 \sin(kx - \omega t + \delta_2) + E_3 \sin(kx - \omega t + \delta_3) + \dots \quad (1019)$$

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*¹¹⁷ 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_{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?

Let's start with the simplest of interference problems: Two Slit Interference.

¹¹⁷Since 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.

13.2: 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 (178), 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.

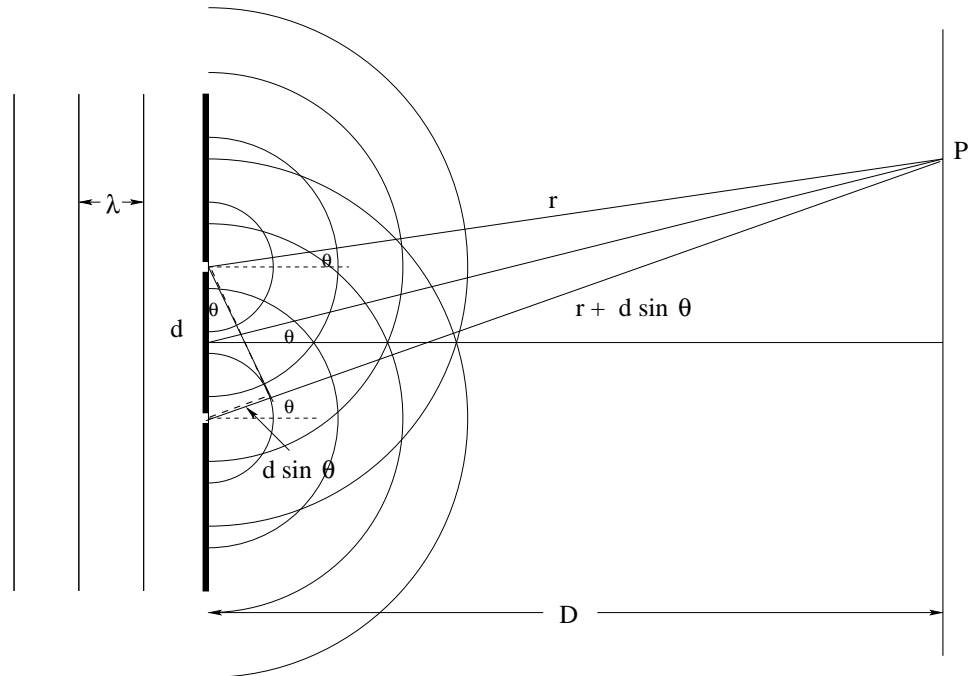


Figure 178: Two narrow slits act as Huygens radiators when incident plane wavefronts fall upon them. Light from the two slits is *coherent* and *in phase* as it leaves the slits, but arrives at P with a phase difference that depends on the path difference.

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:

$$\begin{aligned}
 E_{\text{tot}}(P) &= E_0 \frac{D}{r} \sin(kr - \omega t) + E_0 \frac{D}{r + \Delta r} \sin(kr + k\Delta r - \omega t) \\
 &= E_0 \frac{D}{r} \sin(kr - \omega t) + E_0 \frac{D}{r} \left(1 + \frac{\Delta r}{r}\right)^{-1} \sin(kr + k\Delta r - \omega t) \\
 &= E_0 \frac{D}{r} \sin(kr - \omega t) + E_0 \frac{D}{r} \left(1 - \frac{\Delta r}{r} + \dots\right) \sin(kr + k\Delta r - \omega t) \\
 &= E_0 \frac{D}{r} \sin(kr - \omega t) + E_0 \frac{D}{r} \sin(kr + k\Delta r - \omega t) + \mathcal{O}\left(\frac{\Delta D}{r}\right) \\
 &\approx E_0 \sin(kr - \omega t) + E_0 \sin(kr - \omega t + \delta)
 \end{aligned} \tag{1020}$$

The last step follows because for a small angle θ :

$$r = \frac{D}{\cos(\theta)} \approx \left(\frac{D}{1 - \frac{\theta^2}{2} + \dots}\right) \approx D \left(1 + \frac{\theta^2}{2} + \dots\right) \approx D \tag{1021}$$

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*:

$$\delta = k\Delta r = kd \sin(\theta) = \frac{2\pi d}{\lambda} \sin(\theta) \tag{1022}$$

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 (179) 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 isosceles it is easy to see that the two acute angles must both be $\delta/2$ ¹¹⁸. The total amplitude is thus the sum of the adjacent side lengths of the two right triangles formed by dropping a normal as shown:

$$|E_{\text{tot}}| = 2E_0 \cos(\delta/2) \tag{1023}$$

and the full time dependent electric field is given by:

$$E_{\text{tot}} = 2E_0 \cos(\delta/2) \sin(kr - \omega t + \delta/2) \tag{1024}$$

¹¹⁸The 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 isosceles, 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.

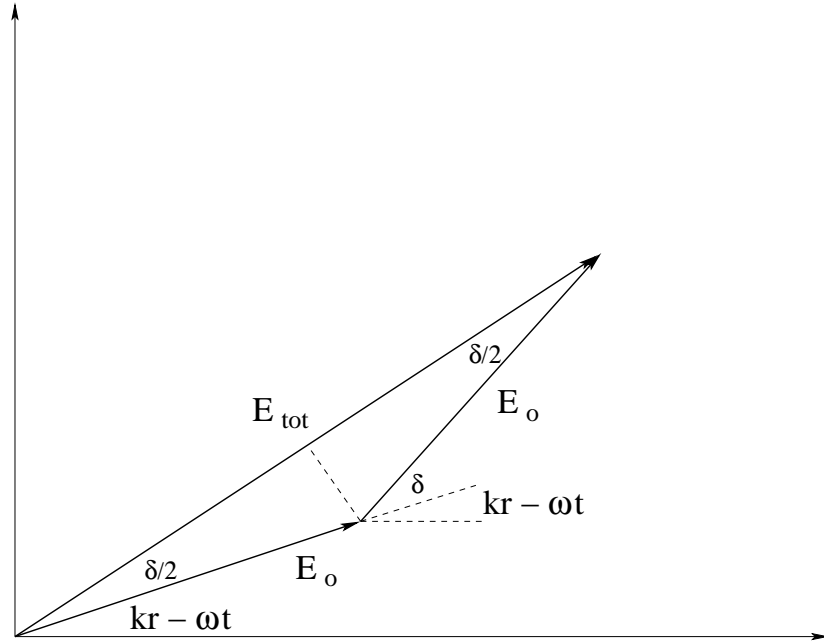


Figure 179: Phasor diagram for the addition of the electric field components of two slits.

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:

$$I_0 = \frac{1}{2\mu_0 c} |E_0|^2 \quad (1025)$$

(from the Poynting vector, as we have seen many times at this point). The total intensity from the pair of slits is therefore:

$$I_{\text{tot}} = 4I_0 \cos^2(\delta/2) \quad (1026)$$

as you should show, filling in the missing steps.

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:

$$\delta/2 = \pm\pi/2, \pm3\pi/2, \pm5\pi/2, \dots \quad (1027)$$

or

$$\delta = \frac{2\pi d}{\lambda} \sin(\theta) = \pm(2m+1)\pi \quad (1028)$$

or the actual angles θ where:

$$d \sin(\theta) = \pm \frac{2m+1}{2} \lambda \quad (1029)$$

The intensity is zero at the minima.

The maxima occur at the angles where:

$$\delta/2 = 0, \pm\pi, \pm2\pi \dots \quad (1030)$$

or

$$\delta/2 = \frac{2\pi d}{2\lambda} \sin(\theta) = m\pi \quad (1031)$$

or the actual angles θ where:

$$d \sin(\theta) = \pm m\lambda \quad (1032)$$

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.

13.3: Interference from Three Narrow Slits

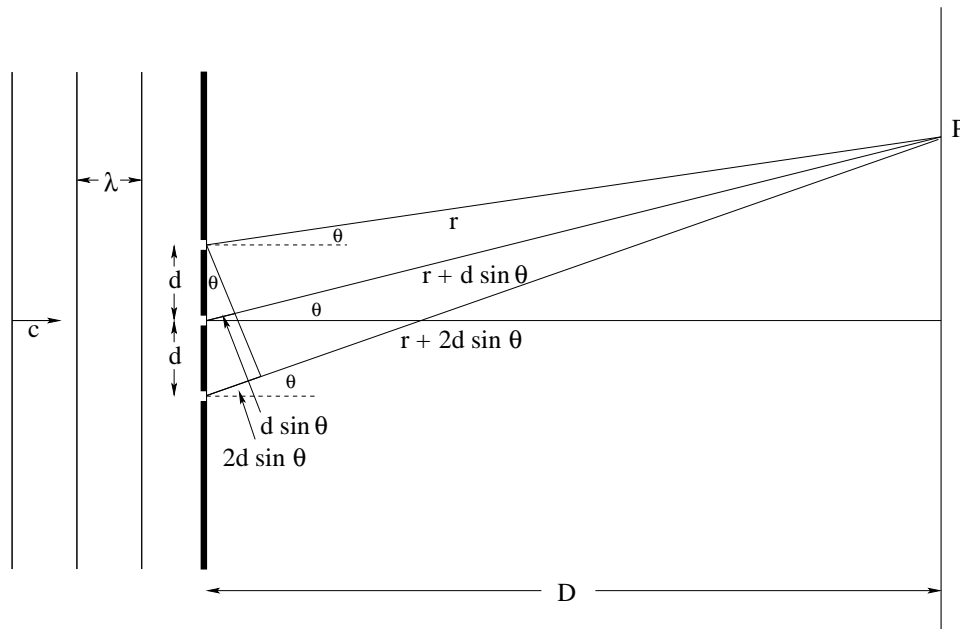


Figure 180: Three narrow slits, equally spaced a distance $d > \lambda$ apart, are illuminated by monochromatic light that is coherent over distances long with respect to both d and λ to produce an interference pattern on a distant screen. Note well that the path difference between any adjacent pair of slits is $d \sin(\theta)$.

In the case of three narrow slits, each separated by the same distance d (illustrated in figure 180), we can follow a more or less identical procedure to find the overall amplitude from a phasor diagram and square it to find the intensity on the screen in terms of the intensity produced by a single slit. We can also *begin* the process of identifying general rules for finding the angle and amplitude (at least approximately) of important features of the interference pattern produced, rules that will work for four, five, or indefinitely many slits. As before we will assume that Fraunhofer conditions hold: the screen is “far” (compared to d and λ) from the slits, and either we will confine our attentions only to angles that are near the center of the screen or we will consider the screen to “wrap around” the slits in the shape of a cylinder so that it is all an equal distance from the central slit¹¹⁹.

¹¹⁹Not that we couldn't explicitly include the effect of r 's gross variation with angle, especially if we programmed

Consider the general phasor diagram in figure ???. We wish to add:

$$E_{\text{tot}} = E_0 \sin(kr - \omega t) + E_0 \sin(kr - \omega t + \delta) + E_0 \sin(kr - \omega t + 2\delta) \quad (1033)$$

with $\delta = kd \sin(\theta)$ is the phase angle produced by the path difference between any two adjacent slits¹²⁰. Examining figure (181) we see that the general result is:

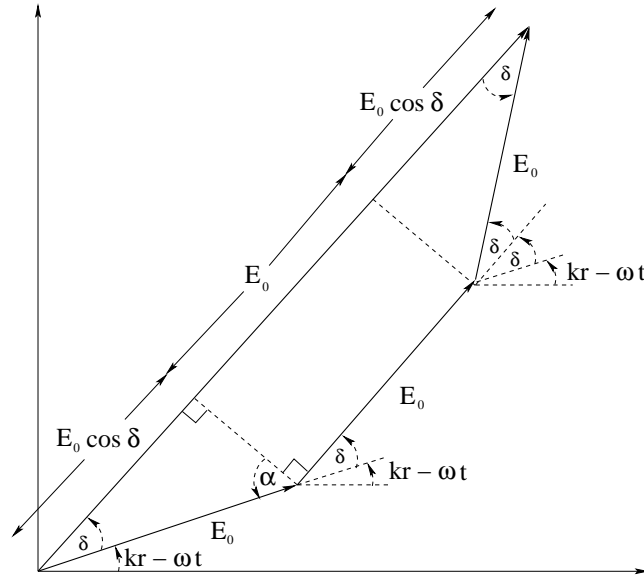


Figure 181: Phasor diagram for general solution for three slits. Note that the amplitude of the sum of the three phasors $E_{\text{tot}} = E_0 + 2E_0 \cos(\delta)$.

$$E_{\text{tot}} = E_0(1 + 2 \cos(\delta)) \quad (1034)$$

and we rather expect that the interference pattern intensity will be:

$$I_{\text{tot}} = \frac{1}{2\mu_0 c} |E_{\text{tot}}|^2 = I_0 (1 + 4 \cos(\delta) + 4 \cos^2(\delta)) \quad (1035)$$

which equals $9I_0$ when $\delta = 0, 2\pi, 4\pi, \dots$ and equals I_0 when $\delta = \pi, 3\pi, 5\pi, \dots$. It *seems* as though it will equal zero for certain values of the phase angle as well, but how can we determine which ones?

To answer this last question and find a more general way of determining the pattern of maxima and minima for 3 slits (and later for more) we turn back to the phasor diagram. Consider the four diagrams drawn in figure (182):

Clearly, we get a principle maximum whenever the three phasors line up (for simplicity the figures are shown at a time that $kr - \omega t = 0$) for a *total* field amplitude of $3E_0$. This obviously occurs when $\delta = 0$, but it can *also* correspond to $\delta = 2\pi, 4\pi, 6\pi, \dots$ – rotating any field phasor through 2π puts it back where it started. We conclude that this arrangement leads to a *maximum* in intensity with $I_p = 9I_0$ called the *principle maxima* of the interference pattern, when the condition:

$$\delta_{\text{principle max}} = \frac{2\pi}{\lambda} d \sin(\theta) = 0, \pm 2\pi, \pm 4\pi, \dots = \pm 2\pi m \quad m = 0, 1, 2, \dots \quad (1036)$$

a computer to do the tedious arithmetic for us, but this course isn't about doing hard arithmetic – seriously, stop laughing – it is about *ideas* and the *idea* of interference can be perfectly well understood and quantitatively analyzed with these simplifications, idealizations, and approximations.

¹²⁰Note well that the angles in the corners of the symmetric trapezoid can be seen to equal δ by reasoning out loud: “ δ plus $\pi/2$ plus α add up to π because they make a straight line. Inside the bottom triangle, α plus $\pi/2$ plus the unknown angle in the corner at the origin add up to π because it is a triangle. Therefore the bottom angle must be δ .”

And you thought high school geometry wasn't good for anything...

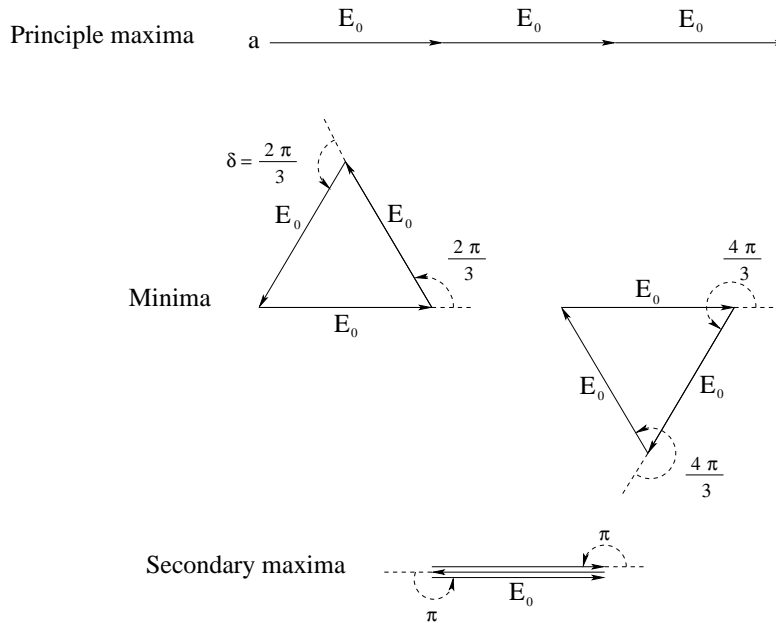


Figure 182: Phasor diagrams illustrating principle maxima, minima, and secondary maxima in the interference pattern. Note that we get minima when the three phasors *close* to get a three-sided polygon or 3-gon (a.k.a. an equilateral triangle in this case). In between the minima we get maxima, but the secondary maxima are much weaker than the principle maxima that occur when all three slits arrive in phase because $d \sin(\theta) = m\lambda$.

If we divide by 2π and multiply by λ , we see that this corresponds to:

$$d \sin(\theta) = \pm m\lambda \quad (1037)$$

just as before for two slits separated by d , so that the angles for principle maxima are:

$$\theta_m^{\text{principle max}} = \frac{\pm m\lambda}{d} \quad (1038)$$

This is *important*: **The location of the principle maxima of N slits is determined by the slit separation d , not by 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 .

Now let's consider the minima. We immediately note that the intensity cannot be negative – I don't 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, which, given three equal field strengths, occurs when the phasors form a *closed, three sided figure*, that is, a *unilateral triangle*¹²¹. The two triangles in the figure above thus represent phase angles that lead to *minima*.

We observe that we close these triangles when:

$$\delta_{\min} = \frac{2\pi}{3} \text{ or } \frac{4\pi}{3} \quad (1039)$$

¹²¹To 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.

or these angles with *any integer multiple of 2π added (or subtracted)*. If we multiply this out and turn it into a rule, it becomes:

$$\begin{aligned}\delta_{\min} = kd \sin(\theta) &= \frac{2\pi}{3}, \frac{4\pi}{3}, \frac{8\pi}{3}, \frac{10\pi}{3}, \frac{14\pi}{3}, \dots \\ \frac{2\pi}{\lambda} d \sin(\theta) &= \frac{2\pi}{3}, \frac{4\pi}{3}, \frac{8\pi}{3}, \frac{10\pi}{3}, \frac{14\pi}{3}, \dots \\ d \sin(\theta) &= \frac{m\lambda}{3} \quad m = \otimes, 1, 2, \otimes, 4, 5, \otimes, 7, 8, \dots\end{aligned}\quad (1040)$$

Note that this is *almost* the integer multiples of $2\pi/3$ (where 3, recall, is the *number of slits* – hmmm, one wonders if this rule *generalizes...*). However, 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*. I remind you of this 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 corresponds to a *secondary maximum*. If we set:

$$\delta_{\text{secondary max}} = \pi, 3\pi, 5\pi, \dots \quad (1041)$$

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:

$$\begin{aligned}\delta_{\text{secondary max}} &= \pi, 3\pi, 5\pi, \dots \\ &= \frac{2\pi m}{2} \quad m = \otimes, 1, \otimes, 3, \otimes, 5, \dots\end{aligned}\quad (1042)$$

which looks like it *might* be a rule involving $2\pi m/(N-1)$ with the usual skip-all- m -that-lead-to-a-multiple-of- 2π constraint.

The expressions for θ_m^{\min} and $\theta_m^{\text{secondary max}}$ are now pretty obvious, and I'll leave you to find them on your own. 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. We'll therefore (in the next section) skip 4 slits and jump right on up to 5 slits, and from there to an arbitrary finite number N of slits. We won't quite prove that the rules we infer (which all work for $N = 2, 3, 5$ in our examples) work for any number of slits, but we *almost* prove it and cleaning up what we do and turning it into a formal proof isn't difficult, just a bit beyond the scope of this course. Unless, perhaps, you are a physics major and need the practice...

13.4: Interference from 4, 5, ... N Narrow Slits

Now let's look at one more particular case (just to be sure) and then generalize the above results. We will not worry (in *this* course) about actually finding the explicit total electric field and squaring it and factoring it to find the intensity for more than two slits, even though I derived the intensity for three just to show you one way that it can be done. Instead we will (continue to) focus on just finding the angles of the principle maxima, the minima, and *approximately* finding the angles of the secondary maxima. In all cases we will be graphically "evaluating":

$$\begin{aligned}E_{\text{tot}} &= E_0 \sin(kr - \omega t) + E_0 \sin(kr - \omega t + \delta) + E_0 \sin(kr - \omega t + 2\delta) + \dots \\ &\quad + E_0 \sin(kr - \omega t + (N-2)\delta) + E_0 \sin(kr - \omega t + (N-1)\delta)\end{aligned}\quad (1043)$$

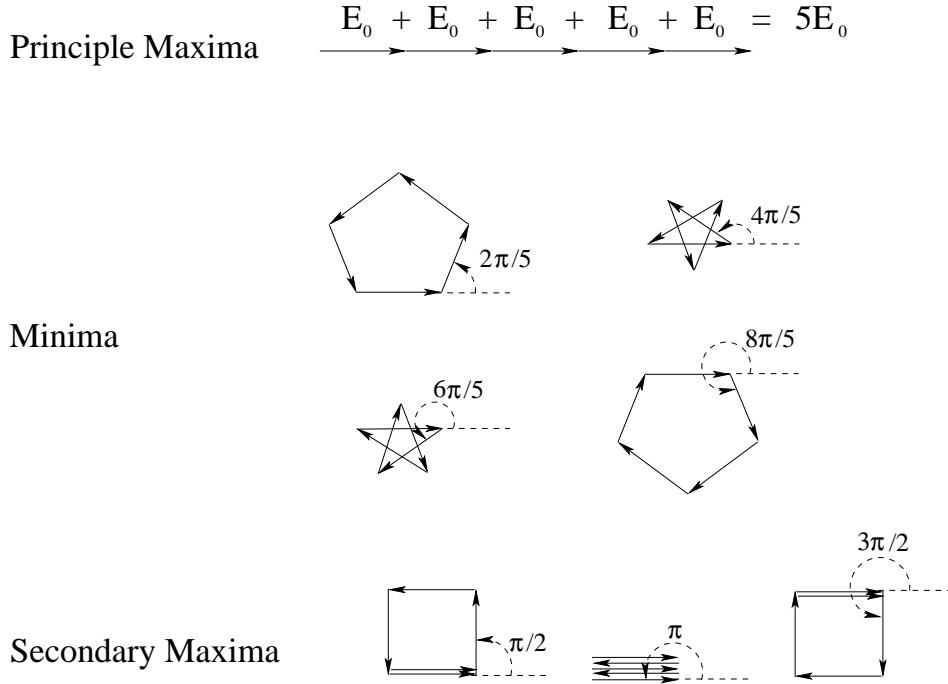


Figure 183: Phasor diagrams principle maxima, minima, and secondary maxima for five slits. The amplitude of the secondary maxima aren't *exactly* E_0 (or equal) and the angles aren't *exactly* at $\delta = 2\pi/(N - 2)$ (for $N = 5$) but this is close enough for an excellent semi-quantitative graph of the intensities (and our heuristic understanding).

where $\delta = kd \sin(\theta)$ is the phase angle produced by the path difference between any **two adjacent slits** in a set of N slits.

To see exactly how the results generalize, let's draw the phasors for one more set of slits, this one with $N = 5$, in figure 183. That should be plenty for us to infer a rule and understand how diffraction gratings (our next subject) and single slit diffraction (the one after that) work.

Note the following features, described in terms of the *general* rules that they represent:

- 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:

$$\delta_{\text{principle max}} = 2\pi m \quad m = 0, 1, 2, 3... \quad (1044)$$

or:

$$\begin{aligned} \delta_{\text{principle max}} = kd \sin(\theta) &= 2\pi m \\ \frac{2\pi}{\lambda} d \sin(\theta) &= 2\pi m \\ d \sin(\theta) &= m\lambda \end{aligned} \quad (1045)$$

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 sense.

Note well that the series doesn't continue indefinitely – the largest m that contributes is one where:

$$\theta_m^{\text{principle max}} = \sin^{-1} \left(\frac{m\lambda}{d} \right) \quad (1046)$$

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:

$$\delta_{\min} = \frac{2\pi m}{5} \quad m = \otimes, 1, 2, 3, 4, \otimes, 6, 7, 8, 9, \otimes, \dots \quad (1047)$$

where I've \otimes 'd out the values $m = 0, 5, 10, \dots$. 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:

$$\delta_{\min} = \frac{2\pi m}{N} \quad \text{for} \quad m = \otimes, 1, 2, \dots, N-1, \otimes, N+1, N+2, \dots, 2N-1, \otimes, 2N+1, \dots \quad (1048)$$

where we have to skip every N th value of m .

Take a moment and verify that this rule works for $N = 2$ and $N = 3$ slits, and derive the related expression for $d \sin(\theta_m)$ and hence θ_m^{\min} for N 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:

$$\delta_{\text{secondary max}} = \frac{2\pi m}{4} = \frac{\pi m}{2} \quad m = \otimes, 1, 2, 3, \otimes, 5, 6, 7, \otimes \dots \quad (1049)$$

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:

$$\delta_{\text{secondary max}} = \frac{2\pi m}{N-1} \quad m = \otimes, 1, 2, \dots, N-2, \otimes, N, N+1 \dots \quad (1050)$$

and so on.

These rules are more than sufficient to allow us to draw a qualitatively correct graph both of the intensity produced by 5 slits and a “generic” graph of “ N ” slits (where of course we have to pick some large but finite number to illustrate).

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) is pretty much *identical to the laws we observe today!*

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 there, then, seems 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 diffraction 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 an intensity pattern, cast on a distant screen, that is the result of part of the light passing through the slit interfering with the rest, a wave interfering with itself. In the next two sections we will therefore apply the concepts we have learned for 2, 3, ..., N slits, beginning with N -slit interference for large N straight up, the *diffraction grating*.

13.5: The Diffraction Grating – Rayleigh’s Criterion for Resolution

Consider now a *diffraction grating* – basically an opaque material with many transparent narrow slits inscribed through the opacity, each separated from its neighbor by a distance d . 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 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:

$$\theta_m^{\max} = \sin^{-1} \left(\frac{m\lambda}{d} \right) \quad (1051)$$

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:

$$\theta_m^{\min} = \sin^{-1} \left(\frac{n\lambda}{Nd} \right) \quad (1052)$$

for the particular values:

$$\begin{aligned} n_1 &= N \pm 1 \\ n_2 &= 2N \pm 1 \\ &\dots \\ n_m &= mN \pm 1 \\ &\dots \end{aligned} \quad (1053)$$

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 m th principle maximum:

$$n_m = N * m - 1, N * m + 1 \quad m = 1, 2, 3... \quad (1054)$$

We now no longer need n_m . We can directly write these angles in terms of m alone as (factoring):

$$\theta_m^{\min} = \sin^{-1} \left(\frac{m\lambda}{d} \pm \frac{\lambda}{Nd} \right) \quad (1055)$$

for each pair of values that bracket the m th 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:

$$\theta_m^{\max} \approx \left(\frac{m\lambda}{d} \right) \quad (1056)$$

and:

$$\theta_m^{\min} \approx \left(\frac{m\lambda}{d} \pm \frac{\lambda}{Nd} \right) = \theta_m^{\max} \pm \frac{\lambda}{Nd} \quad (1057)$$

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:

$$\theta_1(\lambda_1) = \frac{\lambda_1}{d} \quad (1058)$$

while that for λ_2 occurs at:

$$\theta_1(\lambda_2) = \frac{\lambda_2}{d} \quad (1059)$$

These two lines are *separated* in angle by:

$$\Delta\theta_{12} = |\theta_1 - \theta_2| = \frac{\lambda_1 - \lambda_2}{d} \quad (1060)$$

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:

$$\Delta\theta \approx \frac{2\lambda_1}{Nd} \approx \frac{2\lambda_2}{Nd} \quad (1061)$$

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*.

13.5.1: 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.

13.5.2: 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:

$$\theta_m(\lambda_2) = \frac{m\lambda_2}{d} > \frac{m\lambda_1}{d} + \frac{\lambda_1}{Nd} = \theta_m^{\min}(\lambda_1) \quad (1062)$$

or

$$\Delta\lambda_{21} = \frac{m(\lambda_2 - \lambda_1)}{d} > \frac{\lambda_1}{Nd} \quad (1063)$$

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:

$$R = mN = \frac{\lambda}{\Delta\lambda} \quad (1064)$$

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:

$$\Delta\lambda_{\min} = \frac{\lambda}{mN} \quad (1065)$$

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 sweep 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.

13.6: 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 suggests 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.

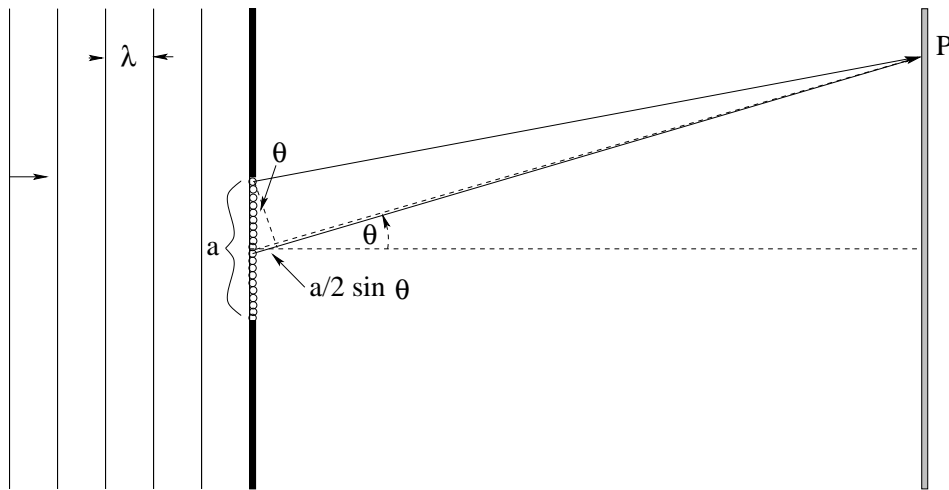


Figure 184: The geometry of single slit diffraction. Waves of some wavelength λ pass through a slit of width a , where a is typically somewhat larger than λ (to get an “interesting” diffraction pattern) and fall upon a screen under Fraunhofer conditions, where the screen is distant compared to a and λ and roughly equidistant from the center of the slit

The geometry of diffraction is straightforward and is represented in figure 184. 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 bother¹²².

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 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.

¹²²We’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 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.

13.7: Diffraction Minima, Heuristic Rule

Consider the two waves emerging from the two Huygens radiators portrayed above in figure 184 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

$$\begin{aligned} \frac{a}{2} \sin(\theta) &= \frac{\lambda}{2}, \text{ or} \\ a \sin(\theta) &= \lambda \end{aligned} \tag{1066}$$

a condition for a *diffraction minimum*.

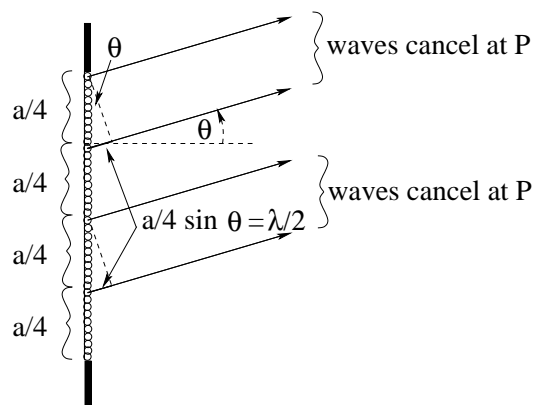


Figure 185: The slit, with the Huygens radiators divided into four equal segments. Light from the two pairs indicated cancels at P when the path difference $\frac{a}{4} \sin(\theta)$ contains a half of a wavelength, for all of the pairs that make up the slit.

Now imagine dividing the strip into fourths, as portrayed in figure 185. 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:

$$a \sin(\theta) = 2\lambda \quad (1067)$$

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 eighths we get $a \sin(\theta) = 4\lambda$. Clearly we can continue indefinitely; the general rule for a minimum is:

$$a \sin(\theta) = m\lambda \quad m = \otimes, 1, 2, 3, \dots \quad (1068)$$

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).

13.8: Exact Solution to Diffraction by a Single Slit

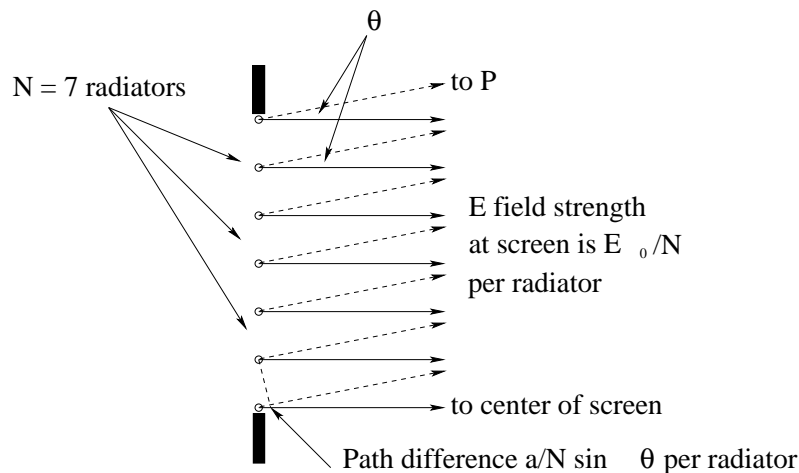


Figure 186: If we split the slit up into N radiators, the field amplitude at the maximum in the center of the screen from *each* radiator is E_0/N , where E_0 is the maximum amplitude from the entire slit there. When we consider the waves emerging at an angle θ directed towards point P , each radiator 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 $N \rightarrow \infty$ and fill in the entire slit with radiators.

In figure 186 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 $N \rightarrow \infty$ so that we *really* cover the slit with a continuum of radiators¹²³ 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 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 $N \rightarrow \infty$ we will make an error of order $1/N$ using the first relation¹²⁴ so we’ll just ignore it and use the first (easier) relation.

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 $N = 7$):

$$\begin{aligned}
 E_{\text{tot}} = & \frac{E_0}{N} \sin(kr - \omega t) + \frac{E_0}{N} \sin(kr - \omega t + \phi/N) \\
 & + \frac{E_0}{N} \sin(kr - \omega t + 2\phi/N) + \frac{E_0}{N} \sin(kr - \omega t + 3\phi/N) \\
 & + \frac{E_0}{N} \sin(kr - \omega t + 4\phi/N) + \frac{E_0}{N} \sin(kr - \omega t + 5\phi/N) \\
 & + \frac{E_0}{N} \sin(kr - \omega t + 6\phi/N)
 \end{aligned} \tag{1069}$$

¹²³... 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...

¹²⁴As you can easily see by doing the binomial expansion of $a/(N-1) = (a/N)(1-1/N)^{-1}$, right...?

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:

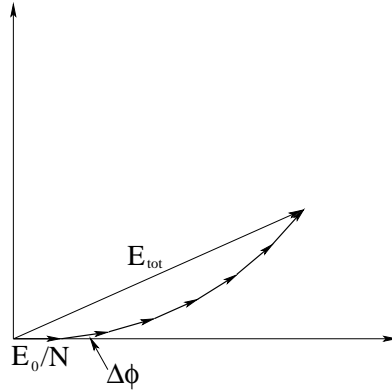


Figure 187: The phasor diagram for $N = 7$ Huygens radiators distributed across a . The amplitude of each radiator is E_0/N , and the phase $\Delta\phi = \phi/N$ accumulates.

The diagram in figure 187 (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 Fraunhofer approximation as¹²⁵:

$$E_{\text{tot}} = \lim_{N \rightarrow \infty} \sum_{i=0}^N \frac{E_0}{N} \sin(kr - \omega t + i\phi/N) \quad (1070)$$

This sort of sum, accumulating infinitesimal chunks of E at infinitesimally different phase angles, is begging to be turned into an integral¹²⁶, 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 188.

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 the bottom has to match the angle

¹²⁵Note that we are still ignoring that extra $\mathcal{O}(N)$ term on the end as there are $N + 1$ terms in the sum.

¹²⁶Ideally 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...

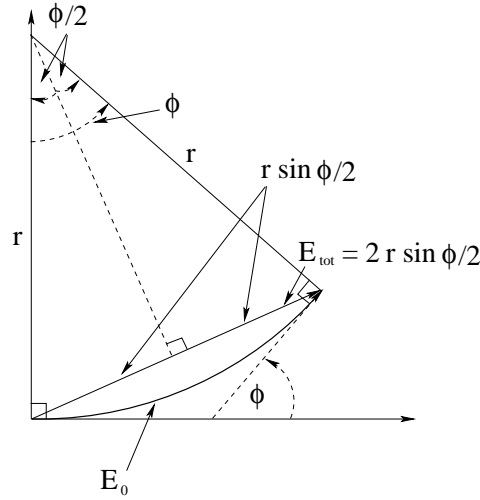


Figure 188: The phasor diagram for $N \rightarrow \infty$ Huygens radiators distributed across a . The “phasor snake” bends smoothly around into a circular arc of length E_0 , where we need to determine the length of the secant that cuts across, E_{tot} .

accumulated between the radii. From this we see that the arc length E_0 can be related to r by:

$$E_0 = r\phi \quad (1071)$$

If we drop a perpendicular bisector (dashed line) from the center of the circular arc to the total field phasor E_{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:

$$E_{\text{tot}} = 2r \sin(\phi/2) \quad (1072)$$

We substitute $r = E_0/\phi$ into this (eliminating r in favor of E_0) to get:

$$E_{\text{tot}} = \frac{2E_0 \sin(\phi/2)}{\phi} = E_0 \left(\frac{\sin(\phi/2)}{\phi/2} \right) \quad (1073)$$

Finally, we go through the usual ritual to convert the field amplitudes to intensities:

$$I_0 = \frac{1}{2\mu_0 c} E_0^2 \quad (1074)$$

so that:

$$I_{\text{tot}} = \frac{1}{2\mu_0 c} E_{\text{tot}}^2 = \frac{1}{2\mu_0 c} E_0^2 \left(\frac{\sin(\phi/2)}{\phi/2} \right)^2 \quad (1075)$$

or

$$I_{\text{tot}}(\theta) = I_0 \left(\frac{\sin(\phi/2)}{\phi/2} \right)^2. \quad (1076)$$

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.

At a glance, this equation has all of the right features. At $\theta = 0$ (and hence $\phi = 0$) we get an intensity of I_0 ¹²⁷. At all the other places where $\sin(\phi/2) = 0$, we get a minimum.

¹²⁷We avoid the problem of “division by zero” calculus-fashion by taking the *limit*

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = \frac{x - x^3/3! + x^5/5! - \dots}{x} = 1 - x^2/3! + x^4/5! - \dots = 1$$

This occurs when:

$$\frac{\phi}{2} = \frac{\pi a}{\lambda} \sin(\theta) = \pi, 2\pi, 3\pi \dots \quad (1077)$$

or when:

$$a \sin(\theta) = m\lambda \quad m = 0, 1, 2, 3, \dots \quad (1078)$$

as before, so our heuristic rule is precisely derived.

We can now also (at least in principle) tackle the *maxima*. We will get a maximum in intensity at the values of ϕ for which:

$$\frac{dI_{\text{tot}}}{d\phi} = 0 \quad (1079)$$

and which aren't the minima (which will also occur, recall, at the zeros in the slope of the intensity). Physics majors and advanced students will enjoy this exercise in calculus, which leads one to the relatively simple result that maxima occur when the transcendental equation¹²⁸

$$\frac{\phi}{2} = \tan\left(\frac{\phi}{2}\right) \quad (1080)$$

is satisfied. If one plots $\phi/2$ and $\tan(\phi/2)$ simultaneously on a single set of axes, the intersections of the two lines are the relevant zeros. As one can see (once one does this) the maxima occur at angles *close to* (and just before) the condition(s):

$$\phi/2 = 0 \text{ (exact)}, 3\pi/2, 5\pi/2, 7\pi/2 \dots \quad (1081)$$

(note well the skipping of $\pi/2$).

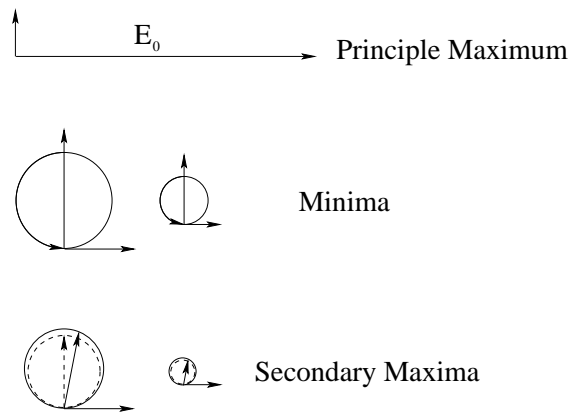


Figure 189: Phasor diagrams representing successive minima and maxima for single slit diffraction.

In figure 189 the principle maximum (of length E_0) is illustrated for angle $\phi = 0$. The next two phasors show the (exact) conditions for minima, where E_0 is wrapped first one time around $\phi = 2\pi$ or twice around $\phi = 4\pi$. Note that the diameter of the circle has to get smaller as one wraps more than once! 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. At $\phi = 3\pi/2$ (dashed circle and arrow) we can gain a tiny bit of length by rolling the circle back to a slightly larger diameter, ditto at $\phi = 5\pi/2$, although both of these figures are probably a bit exaggerated.

It is now time to put it all together with a few examples.

¹²⁸Wikipedia: [http://www.wikipedia.org/wiki/Transcendental Equation](http://www.wikipedia.org/wiki/Transcendental_Equation).

Example 13.8.1: 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:

$$\begin{aligned} a \sin(\theta_m) &= m\lambda \\ 4\lambda \sin(\theta_m) &= m\lambda \\ \sin(\theta_m) &= \frac{m}{4} \\ \theta_m &= \sin^{-1}\left(\frac{m}{4}\right) \end{aligned} \quad (1082)$$

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 6: 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.

13.9: 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:

$$E_{\text{tot}}(\theta) = 2E_0 \cos(\delta/2) \left(\frac{\sin(\phi/2)}{\phi/2} \right) \quad (1083)$$

Following the usual procedure (using the time average Poynting vector and relation between E_0 and B_0) we get the intensity

$$I_{\text{tot}}(\theta) = 4I_0 \cos^2(\delta/2) \left(\frac{\sin(\phi/2)}{\phi/2} \right)^2 \quad (1084)$$

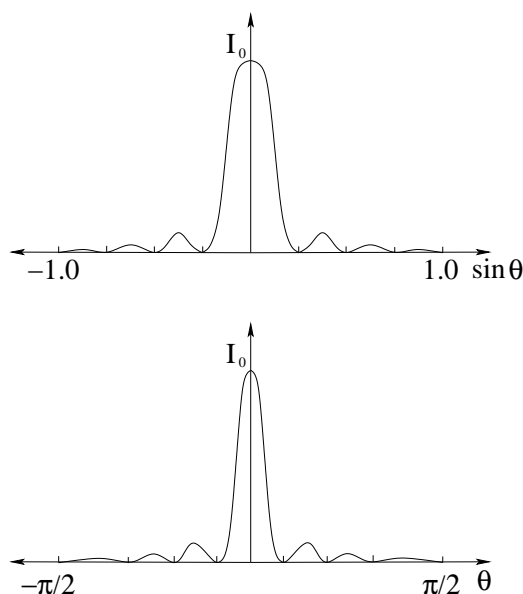


Figure 190: Typical graphs of the diffraction intensity from a single slit of width $a = 4\lambda$. Note the distortion of the horizontal scale by the inverse sine in the lower graph – the top graph is much easier to draw and requires no calculator.

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 13.9.1: 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:

$$\sin(\theta_m) = \frac{m\lambda}{d} = \frac{m}{8} \quad (1085)$$

for maxima and

$$\sin(\theta_m) = \frac{(2m+1)\lambda}{2d} = \frac{2m+1}{16} \quad (1086)$$

for minima. The result is displayed in table 7. Using these numbers we can easily enough construct a combined interference/diffraction pattern, displayed in figure 191. 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_{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 tabulated 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 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!

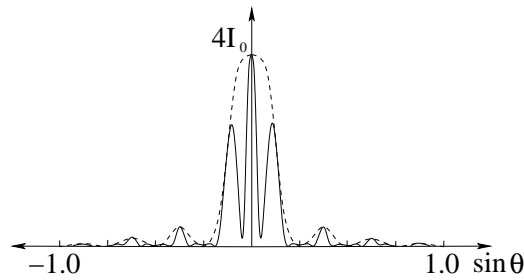


Figure 191: The graph of combined diffraction and interference, for $a = 4\lambda$ (same as before) and $d = 8\lambda$.

13.10: 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 people¹²⁹ 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 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:

$$D \sin(\theta_{\min}) = 1.22\lambda \quad (1087)$$

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

¹²⁹Mostly 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 7: Diffraction minima, interference maxima, and interference minima for a single slit of width $a = 4\lambda$.

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 occurrence 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 patten, 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)¹³⁰.

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

¹³⁰This 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.

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))¹³¹ 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 light¹³²

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:

$$\theta_c \approx \sin(\theta_c) = \frac{1.22\lambda}{D} \quad (1088)$$

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.

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) detail¹³³

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.

¹³¹Wikipedia: 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

¹³²This 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...

¹³³Wikipedia: <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.

13.11: Thin Film Interference

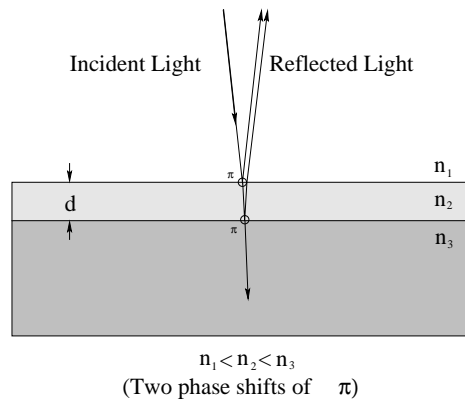


Figure 192: One of the two basic diagrams for thin film interference. The total phase difference in the superposed reflected waves in the case $n_1 < n_2 < n_3$ or $n_3 < n_2 < n_1$ is just $\delta = k'(2d)$, as the phase shifts produced by reflecting off of the two surfaces are either both zero or both (as they are in this case) π , in which case they cancel.

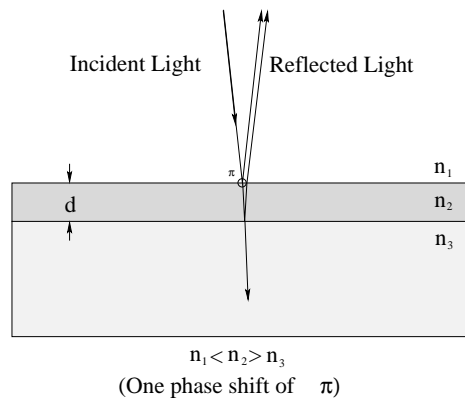


Figure 193: The second of the two basic diagrams for thin film interference. The total phase difference in the superposed reflected waves in the case $n_1 < n_2 > n_3$ or $n_3 < n_2 > n_1$ is $\delta = k'(2d) + \pi$, as there is a phase shift of π produced by reflecting off of the surface of a material with a higher index of refraction only one one of the two surfaces..

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. Fortunately, it is (at this point) quite simple. Here's the idea.

In figures 192 and 193 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:

$$E_{\text{tot}} = E_{12} \sin(kr - \omega t + \delta_{12}) + E_{23} \sin(kr - \omega t + \delta_{23}) \quad (1089)$$

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 off 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).

13.11.1: 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"¹³⁴ on what we have learned so far) to associate with this path difference an additional phase $\delta_{\text{path}} = k(2d)$.

¹³⁴Har, har...

As it turns out, this heuristic guess is *almost* correct! But as the saying goes, “almost” only counts in horseshoes and hand grenades¹³⁵. 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 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:

$$\lambda_2 = \frac{\lambda}{n_2} \quad (1090)$$

where λ is the wavelength of the light in a vacuum. This leads to a wavenumber of:

$$k_2 = \frac{2\pi n_2}{\lambda} \quad (1091)$$

and a phase shift of:

$$\delta_{\text{path}} = k_2(2d) \quad (1092)$$

Basically, the wave that traverses the thin film accumulates phase at the spatial rate of k_2 , not k , k_1 , or k_3 .

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.

13.11.2: 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 ¹³⁶ 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 off of 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 its 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 193 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 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.

¹³⁵...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...

¹³⁶<http://www.phy.duke.edu/~rgb/Class/intro-physics-1.php> Introductory Physics 1

Permutation	δ_{12}	δ_{23}	$ \Delta\delta $
$n_1 < n_2 < n_3$	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 8: Relative phase shift introduced between the wave reflected off of the $n_1 \rightarrow n_2$ interface and the transmitted wave reflected off of the $n_2 \rightarrow n_3$ 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 neighbors or lower, but not in the middle.

13.11.3: 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:

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) = \frac{4\pi n_2 d}{\lambda} \quad (1093)$$

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:

$$2d = m\lambda_2 = m\frac{\lambda}{n_2} \quad m = 0, 1, 2, \dots \quad \text{maxima} \quad (1094)$$

$$2d = \frac{2m+1}{2}\lambda_2 = \frac{(2m+1)}{2}\frac{\lambda}{n_2} \quad m = 0, 1, 2, \dots \quad \text{minima} \quad (1095)$$

Of course, this is only heuristic. The “correct” way to arrive at the same place is to set δ to $0, 2\pi, 4\pi, \dots$ for constructive interference and to $\pi, 3\pi, 5\pi, \dots$ 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.

13.11.4: 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:

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) \pm \pi = \frac{4\pi n_2 d}{\lambda} \pm \pi \quad (1096)$$

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):

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) - \pi = \frac{4\pi n_2 d}{\lambda} - \pi \quad (1097)$$

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:

$$2d = \frac{2m+1}{2}\lambda_2 = \frac{(2m+1)}{2} \frac{\lambda}{n_2} \quad m = 0, 1, 2... \quad \text{maxima} \quad (1098)$$

$$2d = m\lambda_2 = m \frac{\lambda}{n_2} \quad m = 0, 1, 2... \quad \text{minima} \quad (1099)$$

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.

13.11.5: 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 $n_{\text{air}} < n_{\text{oil}} < n_{\text{water}}$.

A second “experiment” – one that is greatly enjoyed by physics students the world over, including very young ones – is to blow soap bubbles¹³⁷. 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

¹³⁷That'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’...

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 alone¹³⁸.

¹³⁸In my online book *Classical Electrodynamics II* I derive the *transmission coefficient*

$$T = \frac{4n_1n_2}{(n_1 + n_2)^2}$$

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 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.

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.

Derive the intensity as a function of θ for the two-slit problem (where the slits are assumed to be $a \ll \lambda$ in width). For $d = 4\lambda$, find the angles where the intensity is maximum and minimum. Sketch the interference pattern from $\theta \in [-\pi/2, \pi/2]$.

Problem 3.

Redo problem 2, but this time assume that the slits have a *finite* width of $a = 3\lambda$ and that $d = 6\lambda$. Determine all of the interference *and* diffraction minima and maxima (the latter can be approximate for diffraction) and sketch a *qualitatively* correct picture of the interference pattern underneath the diffraction envelope.

Problem 4.

There are four permutations of results for thin film interference based on the relative sizes of n_1 , n_2 and n_3 where n_2 is the index of refraction of the thin film itself and the others are the index of refraction of the first (originating medium) and third layers. Derive the condition (relation between t the thickness of the film and λ_0 the wavelength of the incident light in a vacuum) for interference maxima and minima for all four orders. Be sure to circle on your figures the reflections at surfaces that are accompanied by a discrete phase shift of π .

Problem 5.

Draw the phasor diagrams from which the angles at which primary and secondary maxima and minima occur for five small ($a \ll \lambda$) slits separated by a distance d . From these diagrams write the conditions on $\delta = kd \sin \theta$ such that maxima and minima occur. Find the actual angles θ for $d = 4\lambda$, graph the intensity, and compare it to the answer to problem 1 above.

Problem 6.

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 we 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 the determine whether or not Jupiter and its moon are *resolved* by the human eye at this distance.

Problem 7.

Derive the intensity as a function of θ for the single slit problem. For $a = 3\lambda$, find the angles where the intensity is a minimum. Sketch the diffraction pattern from $\theta \in [-\pi/2, \pi/2]$. If you prefer, you can solve for the sines of the angles and sketch the diffraction pattern from $\sin(\theta) \in [-1, 1]$ instead.

Advanced Problem 8.

From your algebraic answer to the previous problem, obtain an expression for the angles where diffraction *maxima* occur. You might find the following useful:

$$\frac{d f^2}{dx} = 2f \frac{df}{dx}$$

which has zeros *both* where $f = 0$ (the minima, except for the one at $\theta = 0$) *and* where $\frac{df}{dx} = 0$ independently. Also recall from the footnote in the text above that:

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$$

and hence is not “undefined”.

Advanced Problem 9.

Derive the expression $R = mN = \frac{\lambda}{\Delta\lambda}$ for resolution for a diffraction grating with N slits of separation d . This proceeds as follows: First use a phasor diagram to determine the angle(s) where the *principle* maxima occur. Then use it to find the angles where the *first minimum* following such a maximum occurs for any given order m . This tells you the *angular half-width* of the maximum for a given λ . Use Raleigh’s criterion for resolution to determine the minimum $\Delta\lambda$ that can be resolved (consider $\lambda' = \lambda + \Delta\lambda$), and verify the expression above.

