Equations du Jour

Introductory Physics II

by

Robert G. Brown

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

Notice

This is a "lecture note" style textbook, designed to support my personal teaching activities at Duke University, in particular teaching its Physics 41/42 series (Introductory Physics for potential physics majors). 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

and will be made available in an inexpensive print version via Lulu press as soon as it is in a sufficiently polished and complete state.

In this way 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 Amazon when that option becomes available, if only to help subsidize me while I continue to write more inexpensive textbooks in physics or other subjects.

Although I no longer use notes to lecture from (having taught the class for decades now, they are hardly necessary) these are 'real' lecture notes and are organized for ease of presentation and ease of learning. They do not try to say every single thing that can be said about each and every topic covered, and are hierarchically organized in a way that directly supports efficient learning.

As a "live" document, these notes have errors great and small, missing figures (that I usually draw from memory in class and will add to the notes themselves as I have time or energy to draw them in a publishable form), 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 without warning as I add content or correct errors. Purchasers of any eventual paper version should be aware of its probable imperfection and be prepared to either live with it or mark up their *own* copies with corrections or additions as need be (in the lecture note spirit) as I do mine. The text has generous margins, is widely spaced, and contains scattered blank pages for students' or instructors' own use to facilitate this.

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

- *Equations du Jour*

  A lecture note style textbook intended to support the teaching of introductory physics.

- *Classical Electrodynamics*

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

Computing Books

- *How to Engineer a Beowulf Cluster*

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

Fiction

- *The Book of Lilith*

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

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

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

- *The Fall of the Dark Brotherhood*

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

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

Poetry

- *Who Shall Sing, When Man is Gone*

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

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

- *Hot Tea!*

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

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

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

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

Both *The Book of Lilith* and *The Fall of the Dark Brotherhood* are also available on Amazon, Barnes and Noble and other online booksellers, and one day from a bookstore near you!

Contents

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.

Note that the *Preliminaries*, *Mathematics* and *Introduction* are not part of the course per se and are not intended to be lectured on. However, it is *strongly suggested that all students read these three chapters right away as their first assignment!* Or, (if you're a student reading these words) you can always decide to read them without it being an assignment, as this book is all about self-actualization in the learning process...

The *Preliminaries* chapter covers not physics but how to *learn* physics (or

anything else). 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.

It is equally strongly suggested that all students *skim read and review the Mathematics chapter* right away, reading it sufficiently carefully that they see what is there so that they can use it as a working reference as they need to while working on the actual course material.

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 at the *beginning* of the book rather than in an appendix that many students never find.
- There are only *twelve chapters*. The book is organized so that it can be sanely taught in a *single college semester* with at *most* a chapter a

week.

- 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. 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.

- One to three problems per chapter are typically "starred" to indicate to the student that these are problems they *must* know how to solve if they wish to do well. These are problems that directly illustrate the relevant concepts and problem solving strategies.

- The textbook is entirely algebraic in its presentation and problem solving requirements – no calculators should be required to solve problems.

This layout provides considerable benefits to both instructor and student. This textbook supports a *top-down* style of learning, where one learns each distinct chapter topic by quickly getting the main points onboard via the summary, then deriving them or exploring them in detail and applying them to example problems, and finally asking the students to use what they have started to learn in highly challenging problems that *cannot* be solved without a deeper level of understanding than that presented in the text.

Part I

Getting Ready

Chapter 1

Preliminaries

1.1 See, Do, *Teach*

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

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

You may find yourself stuck in a course that is taught in a traditional way, by an instructor that lectures, assigns some readings, and maybe on a good day puts on a little dog-and-pony show in the classroom with some audiovisual aids or some demonstrations. The standard expectation in this class is to sit in your chair and watch, passive, taking notes. No real engage-

ment 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 Egyption hieroglyphs without a Rosetta Stone, and the best of luck to you on *that.*

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

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

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

answering them and asking the students questions in turn makes a huge difference. Humans being social creatures, it also makes the class a lot more fun!

In summary, sitting on your ass 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 (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 $\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B})$ 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 $\boldsymbol{B}$ tons of barley with a market value of $\boldsymbol{v}$ 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 $\boldsymbol{v}$ in a magnetic field $\boldsymbol{B}$, and $\boldsymbol{F}$ 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!*

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, and the groups and group activities may be highly structured or freeform. "Studio" methods for teaching physics have even wrapped the lecture itself into a group-structured setting, so *everything* is done in groups, and (probably by making it nearly impossible to be disengaged and sit passively in class waiting for learning to "happen") this approach yields measureable improvements (all things being equal) on at least some objective instruments for measurement of learning.

If you take charge of your own learning, though, you will quickly see that in *any* course, however 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 PDF copy of this book (it's free for instructors, after all, and a click away on the Internet) and point to this page and paragraph containing the following little message from me to them:

> Yo! Teacher! Let's wake up and smell the coffee! Don't prevent your students from doing homework in groups – require it! Make the homework correspondingly more difficult! Give them quite a lot of course credit for doing it well! Construct a recitation or review session where students – in groups – who still cannot get the most difficult problems can get socratic tutorial help *after*

> working hard on the problems on their own! Integrate discussion and deliberately teach to increase *active engagement* (instead of passive wandering attention) in lecture[1]. Then watch as student performance and engagement spirals into the stratosphere compared to what it was before...

Then pray. Some instructors have their egos tied up in things to the point where *they* cannot learn, and then what can you do? If an instructor lets ego or politics obstruct their search for functional methodology, you're screwed anyway, and you might as well just tackle the material on your own. Or heck, maybe their expertise and teaching experience vastly exceeds my own so that their naked words *are* sufficiently golden that any student should be able to learn by just hearing them and doing homework all alone in isolation from any peer-interaction process that might be of use to help them make sense of it all – all data to the contrary.

My own words and lecture – in spite of my 28 years of experience in the classroom, in spite of the fact that it has been twenty years since I actually used lecture notes to teach the course, in spite of the fact I never, ever prepare for recitation because solving the homework problems with the students "cold" *as* a peer member of their groups is useful where copying my privately worked out solutions onto a blackboard for them to passively copy on their papers in turn is useless, in spite of the fact that I wrote this book *similarly* without the use of any outside resource – my words and lecture are *not*. On the other hand, students who work effectively in groups and learn to use this book (and other resources) and do all of the homework "to perfection" might well learn physics quite well without my involvement at all!

Let's understand *why* working in groups has such a dramatic effect on learning. What happens in a group? Well, a lot of *discussion* happens,

[1] Perhaps by using Studio methods, but I've found that in mid-sized classes and smaller (less than around fifty students) one can get very good results without a specially designed classroom by the Chocolate Method – I lecture without notes and offer a piece of chocolate or cheap toy or nifty pencil to any student who catches me making a mistake on the board before I catch it myself, who asks a particularly good question, who looks like they are nodding off to sleep (seriously, chocolate works wonders here, especially when ceremoniously offered). Anything that keeps students *focussed* during lecture by making it into a game, by allowing/encouraging them to speak out without raising their hands, by praising them and rewarding them for engagement makes a huge difference.

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 a group 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 group peers 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 group reaches full understanding.

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 learn-

ing process *with or without the help or permission of your teachers!* A really smart and effective group 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 in the group understands, everybody in the group benefits, *everybody in the group gets the best possible grade on the material.* This process will actually make you (quite literally) more intelligent. You may or may not become smart enough 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 group, 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.

1.2 Other Conditions for Learning

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

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

reading.

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

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

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

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

How did you do?

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

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

People who are downright *brilliant* (and equally lucky) realized in only three seconds (without cheating an extra second or three, you know who you are) that it consisted of the string of odd digits in ascending order followed by the even digits in descending order. Those people probably got it *all*

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

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 measely (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, 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 is *real* – it actually exists, in more than just a metaphorical sense – and if you study long enough and hard enough to obtain a sudden, epiphinaic realization in any topic you are studying, however trivial or complex (like the pattern exposed above) it is quite likely to be accompanied by a purely mental flash of "light". You'll know it when it happens to you, in other words, and it feels *great*.

Unfortunately, the instructor doesn't usually give students a *chance* to experience this in lecture. No sooner is one seemingly random factoid laid out on the table than along comes a new, apparently disconnected one that

pushes it out of place long before we can either memorize it the hard way or make sense out of it so we can remember it with a lot less work. This isn't really anybody's fault, of course; the light bulb is quite unlikely to go off in lecture *just* from lecture no matter *what* you or the lecturer do – it is something that happens to the prepared mind at the end of a process, not something that just fires away every time you hear a new idea.

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

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

A chess *master*, on the other hand, can play umpty games at once, blindfolded, against pitiful fools like myself and when they've finished winning them all they can go back and recontruct *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 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 if this were a different course). So why bother wasting brain energy on even the *easy* way to remember this string when doing so is utterly unimportant to you in the grand scheme of all things?

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

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

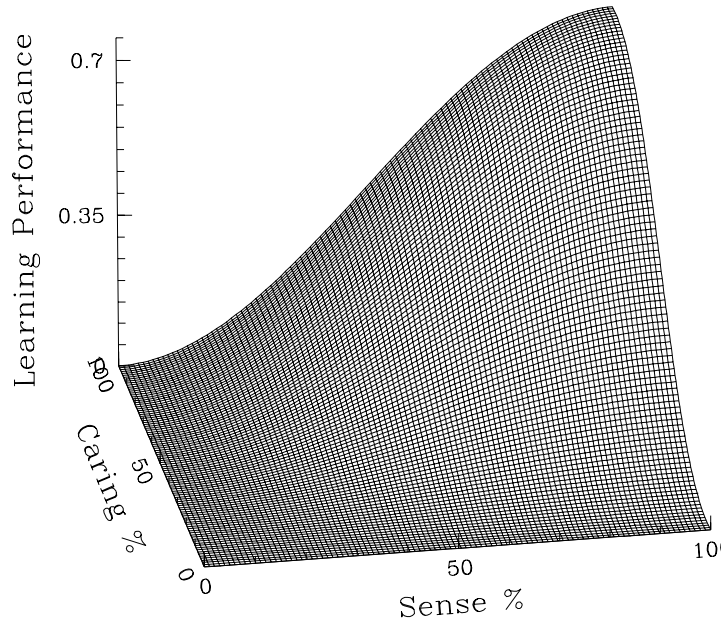


Figure 1.1: Relation between sense, care and learning

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

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

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

Learning is actual work and involves a fair bit of biological stress, just like working out. Your brain needs *food* – it burns a whopping *20-30% of your daily calorie intake all by itself* just living day to day, even more when you are really using it or are somewhat sedentary in your physical habits. Note that your brain runs on pure, energy-rich glucose, so when your blood sugar drops your brain activity drops right along with it. This can happen (paradoxically) because you *just ate a carbohydrate rich meal.* A balanced diet containing foods with a lower glycemic index [3] tends to be harder to digest and provides a longer period of sustained energy for your brain. A daily multivitamin (and various antioxidant supplements such as alpha lipoic acid) can also help maintain your body's energy release mechanisms at the cellular level.

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

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

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

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. Common sense (and yes, experience) will rapidly convince you that learning generally works better if you're not stoned or drunk when you study. Learning works *much* better when you have *time* to learn and haven't put everything off to the last minute. In fact, all of Maslow's hierarchy of needs [4] 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*

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

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 effect may be lost in the statistical noise of the study.

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

1.3 Your Brain and Learning

Your brain is more than just a unique instrument. In some sense it is you. You could imagine having your brain removed from your body and being hooked up to machinary that provided it with sight, sound, and touch in such a way that "you" remain[5]. 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.

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

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 learning methodology that we eventually hope to adopt, let's simply itemize, in no particular order, the various features of the brain [6] 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 has two *cerebral hemispheres* [7] , 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,

[6] Wikipedia: http://www.wikipedia.org/wiki/brain.
[7] Wikipedia: http://www.wikipedia.org/wiki/cerebral_hemisphere.

primate, human. The outermost layer where most actual thinking occurs (in animals that think) is known as the *cerebral cortex*.

- The *cerebral cortex* [8] – 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* [9] . Computer simulated neural networks provide us with insight into how the brain can remember past events and process new information.

- The fundamental operational units of the brain's information processing functionality are called *neurons* [10] . Neurons receive electrochemical signals from other neurons that are transmitted through long fibers called *axons* [11] *Neurotransmitters* [12] 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.

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

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

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

- Intelligence [13] , 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" [14] (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.

[13] Wikipedia: http://www.wikipedia.org/wiki/intelligence.
[14] Wikipedia: http://www.wikipedia.org/wiki/flynn_effect.

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

- Sensory memory [15] 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 $4-5$ objects[16]. 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

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

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

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 megacorporations, you can once again recall the information a day later with only the most cursory of rehearsals. You've taken eighteen "random" objects that were meaningless and could hence be recalled only through the most arduous of rehearsal processes, converted them to six "chunks" of three that can be easily tagged by the brain's existing long term memory (note that you are *not learning* the string FBI, you are building an *association* to the already existing memory of what the string FBI *means*, which is *much easier* for the brain to do), and chunking the chunks into *two* objects.

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

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

- Now we get to the really good stuff. Long term is memory that you form that lasts a long time in human terms. A "long time" can be days, weeks, months, years, or a lifetime. Long term memory is encoded *completely differently* from short term or sensory/immediate memory – it appears to be encoded *semantically* [17] , 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.

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

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

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

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

  Note that this isn't a memory, it is a guess. Guessing is what the mind is designed to do, as it is part of the process by which it "predicts the future" even in the most mundane of ways. When I put ten dollars in my pocket and reach in my pocket for it later, I'm basically guessing,

on the basis of my memory and experience, that I'll find ten dollars there. Maybe my guess is wrong – my pocket could have been picked[18], 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* [19] appears to be partly responsible for both forming spatial maps or visualizations of your environment and also for forming the *cognitive map* that organizes what you know and transforms short term memory into long term memory, and it appears to do its job (as noted above) *in your sleep.* Sleep deprivation *prevents the formation of long term memory.* Being rendered unconscious for a long period often produces *short term amnesia* as the brain loses short term memory before it gets put into long term memory. The hippocampus shows evidence of plasticity – taxi drivers who have to learn to navigate large cities actually have larger than normal hippocampi, with a size proportional to the length of time they've been driving. This suggests (once again) that it is possible to *deliberately increase the capacity* of your *own* hippocampus through the exercise of its functions, and consequently *increase your ability to store and retrieve information*, which is an important component (although not the only component) of intelligence!

- Memory is improved by *increasing the supply of oxygen to the brain*, which is best accomplished by *exercise.* Unsurprisingly. Indeed, as

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

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

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* [20] 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,

[20] Wikipedia: http://www.wikipedia.org/wiki/amygdala.

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* to optimize learning, in so many ways.

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

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

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

1.4 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:

1. Review the whole (typically textbooks and/or notes)

2. Work on the parts (do homework, use it 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 didn't "get it" from one lecture. There was too much material. If you were *lucky* and well prepared and blessed with a good instructor, perhaps you grasped *some* of it for a *moment* (and if your instructor was poor or you were particularly poorly prepared you may not have managed even that) but what you did 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 derviations connecting them to other things you know, a topical association with other things that you know. They tell a *story*, or part of a story, and you need to know that story in *broad* terms, not try to memorize it word for word.

Reviewing the material should be done in layers, skimming the textbook and your notes, creating a *new* set of notes out of the text in combination

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

You especially do *not* have to work on *memorizing* the content. In fact, it is *not* desireable to try to memorize content at this point – you want the big picture *first* so that facts have a place to live in your brain. If you build them a house, they'll move right in without a fuss, where if you try to grasp them one at a time with no place to put them, they'll (metaphorically) slip away again as fast as you try to take up the next one. Let's understand this a bit.

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

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

Now let's understand the second part of doing homework – working prob-

lems. 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 them 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*[21]. The left brain tries to

[21]Note that this description is at least partly metaphor, for while there is some hemi-

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

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

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

---

spherical 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.

1.4.1 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.

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 3-6 hours/week on homework that is expected of you in any course of this level of difficulty!

This ends our discussion of course preliminaries (for nearly *any* course you might take) and it is time to get on with actual material.

The next chapter is on mathematics. It is *not* actually a part of this text; it is a reference of sorts for you to use to refresh your memory of (or learn, as the case may be) things that you *need to know* to make learning the physics *easy*. I suggest quickly reviewing it so you know what is there, then coming *back* to this chapter for help as you need it when working through the physics. Periodically skim it again to refresh it in your mind and build even better associative maps to what is there, and gradually, painlessly, you can build up the critical skills as you work through something else entirely, without even realizing that this is what you are doing.

Chapter 2

Mathematics

This isn't really a math textbook, but math is an extremely important part of physics. Physics textbooks usually at least attempt to include math support for key ideas, reviewing e.g. how to do a cross product. The problem with this is that this topical review tends to be scattered throughout the text or collected in an appendix that students rarely find when they most need it (either way).

I don't really *like* either of these solutions. My own solution is eventually going to be to write a short lecture-note style *math* textbook that contains just precisely what is needed in order to really get going with physics at least through the undergraduate level, *including* stuff needed in the *introductory* classes one takes as a freshman. Most mathematical physics or physical mathematics books concentrate on differential equations or really abstract stuff like group theory. Most intro physics students struggle, on the other hand, with *simple* things like decomposing vectors into components and adding them componentwise, with the quadratic formula, with complex numbers, with simple calculus techniques. Until these things are mastered, differential equations are just a cruel joke.

Math texts tend to be useless for this kind of thing, alas. One would need three or four of them – one for vectors, one for calculus, one for algebra, one for complex numbers. It is rare to find a single book that treats all of this and does so *simply* and without giving the student a dozen examples or exercises per equation or relation covered in the book. What is needed is a *comprehensive review* of material that is shallow and fast enough to let a

student quickly recall it if they've seen it before well enough to use, yet deep and complete enough that they can get to where they can *work* with the math even if they have *not* had a full course in it, or if they can't remember three words about e.g. complex variables from the two weeks three years ago when they covered them.

In the meantime (until I complete this fairly monumental process of splitting off a whole other book on intro math for physics) I'm putting a math review chapter *first* in the book, right here where you are reading these words. I recommend *skimming* it to learn what it contains, then making a slightly slower pass to review it, then go ahead and move on the the physics and come *back* anytime you are stumped by not remembering how to integrate something like (for example):

$$\int_0^\infty x^2 e^{-ax} dx \tag{2.1}$$

Here are some of the things you should be able to find help for in this chapter:

- **Numbers**

  Integers, real numbers, complex numbers, prime numbers, important numbers, the algebraic representation of numbers. Physics is all about numbers.

- **Algebra**

  Algebra is the symbolic manipulation of numbers according to certain rules to (for example) solve for a particular desired physical quantity in terms of others. We also review various well-known functions and certain expansions.

- **Coordinate Systems and Vectors**

  Cartesian, Cylindrical and Spherical coordinate systems in 2 and 3 dimensions, vectors, vector addition, subtraction, inner (dot) product of vectors, outer (cross) product of vectors.

- **Trigonometric Functions and Complex Exponentials**

  There is a beautiful relationship between the complex numbers and trig functions such as sine, cosine and tangent. This relationship is

encoded in the "complex exponential" $e^{i\theta}$, which turns out to be a *very* important and useful relationship. We review this in a way that hopefully will make working with these complex numbers and trig functions both easy.

- **Differentiation**

  We quickly review what differentiation *is*, and then present, sometimes with a quick proof, a table of derivatives of functions that you should know to make learning physics at this level straightforward.

- **Integration**

  Integration is basically antidifferentiation or summation. Since many physical relations involve summing, or integrating, over extended distributions of mass, of charge, of current, of fields, we present a table of integrals (some of them worked out for you in detail so you can see how it goes).

2.1 Numbers

2.1.1 Natural, or Counting Numbers

This is the set of numbers [1] :

$$1, 2, 3, 4 \ldots$$

that is pretty much the first piece of mathematics any student learns. They are used to *count*, initially to count things, concrete objects such as pennies or marbles. This is in some respects surprising, since pennies and marbles are never really *identical*. In physics, however, one encounters particles that *are* – electrons, for example, differ only in their position.

The natural numbers are usually defined along with a set of operations known as *arithmetic* [2] . The well-known operations of arithmetic are addition, subtraction, multiplication, and division. One rapidly sees that the set of natural numbers is not *closed* with respect to them.

[1] Wikipedia: http://www.wikipedia.org/wiki/number.
[2] Wikipedia: http://www.wikipedia.org/wiki/arithmetic.

Natural numbers greater than 1 in general can be factored into a representation in prime numbers. For example:

$$45 = 2^0 3^2 5^1 7^0 ... \tag{2.2}$$

or

$$56 = 2^3 3^0 5^0 7^1 11^0 ... \tag{2.3}$$

2.1.2 Infinity

It is easy to see that there is no largest natural number. Suppose there was one, call it L. Now add one to it, forming $M = L + 1$. We know that $L + 1 = M > L$, contradicting our assertion that L was the largest. This *lack* of a largest object, lack of a boundary, lack of termination in series, is of enormous importance in mathematics and physics. If there is no largest number, if there is no "edge" to space or time, then it in some sense they run on *forever*, without termination.

Still, there are a number of properties of numbers that we can only understand if we *imagine* a very large number used as a boundary or limit in some computation, and then let that number mentally increase without bound. *Note well* that this is a mental trick, encoding the observation that there is no largest number and so we can increase a number parameter without bound, no more. However, we use this mental trick all of the time – it becomes a way for our finite minds to encompass the idea of unboundedness. To facilitate this process, we invent a *symbol* for this unreachable limit to the counting process and give it a name.

We call this unboundedness *infinity* [3] – the lack of a finite boundary – and give it the symbol ∞ in mathematics.

In *many contexts* we will treat ∞ as a number in all of the number systems mentioned below. We will talk blithely about "infinite numbers of digits" in number representations, which means that the digits simply keep on going without bound. However, it is *very important* to bear in mind that ∞ is *not* a number, it is a *concept*. Or at the very least, it is a highly *special* number, one that doesn't satisfy the axioms or participate in the

[3] Wikipedia: http://www.wikipedia.org/wiki/Infinity.

usual operations of ordinary arithmetic. For example:

$$\begin{aligned} \infty + \infty &= \infty && (2.4)\\ \infty + N &= \infty && (2.5)\\ \infty - \infty &= \text{undefined} && (2.6)\\ \infty * N &= \infty && (2.7) \end{aligned}$$

These are certainly "odd" rules compared to ordinary arithmetic!

For a bit longer than a century now (since Cantor organized set theory and discussed the various ways sets could become infinite and set theory was subsequently axiomatized) there has been an *axiom of infinity* in mathematics postulating its formal existence as a "number" with these and other odd properties.

Our principle use for infinity will be as a limit in calculus and in series expansion. We will use infinity to name the *process* of taking a small quantity and making it "infinitely small" (but nonzero) – the idea of the *infinitesimal*, or the complementary operation of taking a large (finite) quantity (such as a limit in a finite sum) and making it "infinitely large". These operations do not always make arithmetical sense, but when they do they are *extremely valuable* as they are at the heart of both series expansions and calculus.

2.1.3 Integers

To achieve closure in addition, subtraction, and multiplication one introduces negative whole numbers and zero to construct the set of *integers*. Today we take these things for granted, but in fact the idea of negative numbers in particular is quite recent. Although they were *used* earlier, mathematicians only accepted the idea that negative numbers were legitimate numbers by the latter 19th century! After all, if you are counting cows, how can you add negative cows to an already empty field? Numbers were thought of as being concrete properties of *things*, tools for bookkeeping, rather than strictly abstract entities about which one could axiomatically reason until well into the Enlightenment[4].

[4]Interested readers might want to look at Morris Kline's *Mathematics: The Loss of Certainty*, a book that tells the rather exciting story of the development of mathematical reasoning from the Greeks to the present, in particular the discovery that mathematical

In physics, integers or natural numbers are often represented by the letters i, j, k, l, m, n, although of course in algebra one *does* have a range of choice in letters used, and some of these symbols are "overloaded" (used for more than one thing) in different formulas.

Integers can in general also be factored into primes, but problems begin to emerge when one does this. First, negative integers will always carry a factor of -1 times the prime factorization of its absolute value. But the introduction of a form of "1" into the factorization means that one has to deal with the fact that $-1 * -1 = 1$ and $1 * -1 = -1$. This possibility of permuting negative factors through all of the positive and negative halves of the integers has to be generally ignored because there is a complete symmetry between the positive and negative half-number line; one simply appends a single -1 to the prime factorization to serve as a reminder of the sign. Second, 0 times anything is 0, so it (and the number ± 1) are generally excluded from the factorization process.

Integer arithmetic is associative, commutative, is closed under addition/subtraction and multiplication, and has lots of nice properties you can learn about on e.g. Wikipedia. However, it is still not closed under division! If one divides two integers, one gets a number that is not, in general, an integer!

This forming of the *ratio* between two integer or natural number quantities leads to the next logical extension of our system of numbers: The rationals.

2.1.4 Rational Numbers

If one takes two integers a and b and divides a by b to form $\frac{a}{b}$, the result will often *not* be an integer. For example, $1/2$ is not an integer, nor is $1/3, 1/4, 1/5...$, nor $2/3, 4/(-7) = -4/7, 129/37$ and so on. These numbers are all the *ratios* of two integers and are hence called *rational numbers* [5] .

Rational numbers when expressed in a base[6] e.g. base 10 have an inter-

reasoning does not lead to "pure knowledge", *a priori* truth, but rather to contingent knowledge that may or may not apply to or be relevant to the real world.

[5]Wikipedia: http://www.wikipedia.org/wiki/rational number.

[6]The *base* of a number is the range over which each digit position in the number

esting property. Dividing one out produces a finite number of non-repeating digits, followed by a finite sequence of digits that repeats cyclically forever. For example:

$$\frac{1}{3} = 0.3333... \tag{2.8}$$

or

$$\frac{11}{7} = 1.571428571428571428... \tag{2.9}$$

Note that finite precision decimal numbers are precisely those that are terminated with an infinite string of the digit 0. If we keep numbers only to the hundredths place, e.g. 4.17, -17.01, 3.14, the assumption is that all the rest of the digits in the rational number are 0 – 3.14000...

It may not be the case that those digits really *are* zero. We will often be multiplying by $1/3 \approx 0.33$ to get an approximate answer to all of the precision we need in a problem. In any event, we generally *cannot* handle an infinite number of digits, repeating or not, in our arithmetical operations, so truncated, base two or base ten, rational numbers are the special class of numbers over which we do much of our arithmetic, whether it be done with paper and pencil, slide rule, calculator, or computer.

If all rational numbers have digit strings that eventually cyclically repeat, what about all numbers whose digit strings do *not* cyclically repeat? These numbers are *not* rational.

2.1.5 Irrational Numbers

An irrational number [7] is one that *cannot be written* as a ratio of two integers e.g. a/b. It is not immediately obvious that numbers like this exist at all. When rational numbers were discovered (or invented, as you prefer) by the Pythagoreans, they were thought to have nearly mystical properties – the Pythagoreans quite literally worshipped numbers and thought that everything in the Universe could be understood in terms of the ratios of natural numbers. Then *Hippasus*, one of their members, demonstrated that

cycles. We generally work and think in base ten because our ten fingers are amount the first things we count! Hence *digit*, which refers to a positional number *or* a finger or toe. However, base two (binary), base eight (octal) and base sixteen (hexadecimal) are all useful in computation, if not physics.

[7] Wikipedia: http://www.wikipedia.org/wiki/irrational number.

for an isoceles right triangle, if one assumes that the hypotenuse and arm are commensurable (one can be expressed as an integer ratio of the other) that the hypotenuse had to be even, but the legs had to be both even and odd, a contradiction. Consequently, it was certain that they could *not* be placed in a commensurable ratio – the lengths are related by an *irrational* number.

According to the (possibly apocryphal) story, Hippasus made this discovery on a long sea voyage he was making, accompanied by a group of fellow Pythagoreans, and they were so annoyed at his *blasphemous* discovery that their religious beliefs in the rationality of the Universe (so to speak) were false that they *threw him overboard* to drown! Folks took their mathematics quite seriously, back then...

As we've seen, all digital representation of finite precision or digital representations where the digits eventually cycle correspond to rational numbers. Consequently its digits in a decimal representation of an irrational number *never* reach a point where they cyclically repeat or truncate (are terminated by an infinite sequence of 0's).

Many numbers that are of great importance in physics, especially $e = 2.718281828...$ and $\pi = 3.141592654...$ are irrational, and we'll spend some time discussing both of them below. When working in coordinate systems, many of the trigonometric ratios for "simple" right triangles (such as an isoceles right triangle) involve numbers such as $\sqrt{2}$, which are also irrational – this was the basis for the earliest proofs of the existence of irrational numbers, and $\sqrt{2}$ was arguably the first irrational number discovered.

Whenever we compute a number answer we *must* use rational numbers to do it, most generally a finite-precision decimal representation. For example, 3.14159 may *look* like π, an irrational number, but it is really $\frac{314159}{100000}$, a rational number that *approximates* π to six significant figures.

Because we *cannot* precisely represent them in digital form, in physics (and mathematics and other disciplines where precision matters) we will often carry important irrationals along with us in computations as *symbols* and only evaluate them numerically at the end. It is important to do this because we work quite often with functions that yield a rational number or even an integer when an irrational number is used as an argument, e.g. $\cos(\pi) = -1$. If we did finite-precision arithmetic prematurely (on computer

or calculator) we might well end up with an *approximation* of -1, such as -0.999998 and could not be sure if it was *supposed* to be -1 or really was supposed to be a bit less.

There are lots of nifty truths regarding irrational and irrational numbers. For example, in between any two rational numbers lie an *infinite* number of *irrational* numbers. This is a "bigger infinity" [8] than *just* the countably infinite number of integers or rational numbers, which actually has some important consequences in physics – it is one of the origins of the theory of deterministic chaos.

2.1.6 Real Numbers

The union of the irrational and rational numbers forms the *real number line*. [9] Real numbers are of great importance in physics. They are closed under the arithmetical operations of addition, subtraction, multiplication and division, where one must exclude only division by zero. Real exponential functions such as a^b or e^x (where a, b, e, x are all presumed to be real) will have real values, as will algebraic functions such as $(a + b)^n$ where n is an integer.

However, as before we can discover arithmetical operations such as the square root operation that lead to problems with closure. For positive real arguments $x \geq 0$, $y = \sqrt{x}$ is real, but probably irrational (irrational for most possible values of x). But what happens when we try to form the square root of negative real numbers? In fact, what happens when we try to form the square root of -1?

This is a bit of a problem. All real numbers, squared, are positive. There *is* no real number that can be squared to make -1. All we can do is *imagine* such a number, and then make our system of numbers bigger to accomodate it. This process leads us to the *imaginary* unit i such that $i^2 = -1$, and thereby to numbers with both real and imaginary parts: Complex numbers.

[8] Wikipedia: http://www.wikipedia.org/wiki/infinity.
[9] Wikipedia: http://www.wikipedia.org/wiki/real line.

2.1.7 Complex Numbers

At this point you should begin to have the feeling that this process of generating supersets of the numbers we already have figured out will never end. You would be right, and some of the extensions (ones we will not cover here) are actually very useful in more advanced physics. However, we have a finite amount of time to *review* numbers here, and complex numbers are the most we will need in *this* course (or even "most" undergraduate physics courses even at a somewhat more advanced level). They are important enough that we'll spend a whole section discussing them below; for the moment we'll just define them.

We start with the unit imaginary number [10] , i. You *might* be familiar with the *naive* definition of i as the square root of -1:

$$i = +\sqrt{-1} \tag{2.10}$$

This definition is common but slightly unfortunate. If we adopt it, we have to be careful *using* this definition in algebra or we will end up proving any of the many variants of the following:

$$-1 = i \cdot i = \sqrt{-1} \cdot \sqrt{-1} = \sqrt{-1 \cdot -1} = \sqrt{1} = 1 \tag{2.11}$$

Oops.

A better definition for i that it is just the algebraic number such that:

$$i^2 = -1 \tag{2.12}$$

and to leave the square root bit out. Thus we have the following cycle:

$$\begin{aligned} i^0 &= 1 \\ i^1 &= i \\ i^2 &= -1 \\ i^3 &= (i^2)i = -1 \cdot i = -i \\ i^4 &= (i^2)(i^2) = -1 \cdot -1 = 1 \\ i^5 &= (i^4)i = i \\ &... \end{aligned} \tag{2.13}$$

[10] Wikipedia: http://www.wikipedia.org/wiki/imaginary unit.

where we can use these rules to do the following sort of simplification:

$$+\sqrt{-\pi b} = +\sqrt{i^2 \pi b} = +i\sqrt{\pi b} \tag{2.14}$$

but where we never actually write $i = \sqrt{-1}$.

We can make all the purely imaginary numbers by simply scaling i with a real number. For example, $14i$ is a purely imaginary number of magnitude 14. $i\pi$ is a purely imaginary number of magnitude π. All the purely imaginary numbers therefore form an *imaginary line* that is basically the real line, times i.

With this definition, we can define an arbitrary complex number z as the sum of an arbitrary real number plus an arbitrary imaginary number:

$$z = x + iy \tag{2.15}$$

where x and y are both real numbers. It can be shown that the roots of any polynomial function can always be written as complex numbers, making complex numbers of great importance in physics. However, their *real* power in physics comes from their relation to exponential functions and trigonometric functions.

Complex numbers (like real numbers) form a *division algebra* [11] – that is, they are closed under addition, subtraction, multiplication, *and division.* Division algebras permit the factorization of expressions, something that is obviously very important if you wish to algebraically solve for quantities.

Hmmmm, seems like we ought to look at this "algebra" thing. Just what *is* an algebra? How does algebra work?

2.2 Algebra

Algebra [12] is a *reasoning process* that is one of the fundamental cornerstones of mathematical reasoning. As far as we are concerned, it consists of two things:

- Representing *numbers* of one sort or another (where we could without loss of generality assume that they are *complex* numbers, since

[11] Wikipedia: http://www.wikipedia.org/wiki/division algebra.
[12] Wikipedia: http://www.wikipedia.org/wiki/algebra.

real numbers are complex, rational and irrational numbers are real, integers are rational, and natural numbers are integer) with *symbols*. In physics this representation isn't only a matter of knowns and unknowns – we will often use algebraic symbols for numbers we know or for parameters in problems even when their value is actually given as part of the problem. In fact, with only a relatively few exceptions, we will prefer to use symbols as much as we can to permit our algebraic manipulations to eliminate as much eventual *arithmetic* (computation involving actual numbers) as possible.

- Performing a sequence of *algebraic transformations* of a set of equations or inequalities to convert it from one form to another (desired) form. These transformations are generally based on the set of arithmetic operations defined (and permitted!) over the field(s) of the number type(s) being manipulated.

That's it.

Note well that it isn't always a matter of solving for some unknown variable. Algebra is just as often used to derive relations and hence gain insight into a system being studied. Algebra is in some sense the *language* of physics.

The transformations of algebra applied to equalities (the most common case) can be summarized as follows (non-exhaustively). If one is given one or more equations involving a set of variables $a, b, c, ...x, y, z$ one can:

1. Add any scalar number or well defined and consistent symbol to both sides of any equation. Note that in physics problems, symbols carry units and it is necessary to add only symbols that *have the same units* as we cannot, for example, add seconds to to kilograms and end up with a result that makes any sense!

2. Subtract any scalar number or consistent symbol ditto. This isn't really a separate rule, as subtraction is just adding a negative quantity.

3. Multiplying both sides of an equation by any scalar number or consistent symbol. In physics one *can* multiply symbols with different units, such an equation with (net) units of meters times symbols given in seconds.

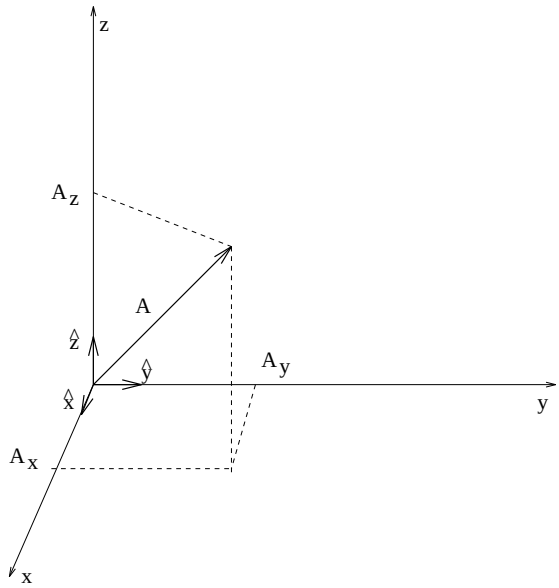
4. Dividing both sides of an equation ditto, save that one has to be careful when performing symbolic divisions to avoid points where division is not permitted or defined (e.g. dividing by zero or a variable that might take on the value of zero). Note that dividing one unit by another in physics is also permitted, so that one can sensibly divide length in meters by time in seconds.

5. Taking both sides of an equation to any power. Again some care must be exercised, especially if the equation can take on negative or complex values or has any sort of domain restrictions. For fractional powers, one may well have to specify the *branch* of the result (which of many possible roots one intends to use) as well.

6. Placing the two sides of any equality into *almost* any functional or algebraic form, either given or known, as if they are variables of that function. Here there are some serious caveats in both math and physics. In physics, the most important one is that if the functional form has a power-series expansion then the equality one substitutes in must be *dimensionless*. This is easy to understand. Supposed I know that x is a length in meters. I *could* try to form the exponential of x: e^x, but if I expand this expression, $e^x = 1 + x + x^2/2! + ...$ which is *nonsense!* How can I add meters to meters-squared? I can only exponentiate x if it is dimensionless. In mathematics one has to worry about the domain and range. Suppose I have the relation $y = 2 + x^2$ where x is a real (dimensionless) expression, and I wish to take the $\cos^{-1}$ of both sides. Well, the *range* of cosine is only -1 to 1, and my function y is clearly strictly larger than 2 and cannot have an inverse cosine! This is obviously a powerful, but *dangerous* tool.

2.3 Coordinate Systems, Points, Vectors

2.4 Review of Vectors



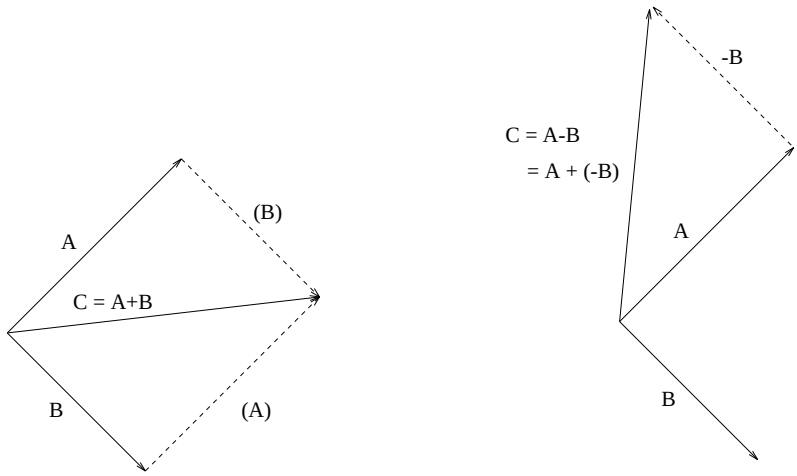
(2.16)

Most motion is not along a straight line. If fact, almost no motion is along a line. We therefore need to be able to describe motion along *multiple dimensions* (usually 2 or 3). That is, we need to be able to consider and evaluate *vector* trajectories, velocities, and accelerations. To do this, we must first learn about what vectors are, how to add, subtract or decompose a given vector in its cartesian coordinates (or equivalently how to convert between the cartesian, polar/cylindrical, and spherical coordinate systems), and what scalars are. We will also learn a couple of products that can be constructed from vectors.

A bf vector in a coordinate system is a directed line between two points. It has **magnitude** and **direction**. Once we define a coordinate origin, each particle in a system has a **position vector** (e.g. – $\vec{A}$) associated with its location in space drawn from the origin to the physical coordinates of the particle (e.g. – (A_x, A_y, A_z)):

$$\vec{A} = A_x\hat{x} + A_y\hat{y} + A_z\hat{z} \tag{2.17}$$

2.4.1 Coordinate Systems and Vectors



The position vectors clearly depend on the choice of coordinate origin. However, the **difference vector** or **displacement vector** between two position vectors does **not** depend on the coordinate origin. To see this, let us consider the **addition** of two vectors:

$$\vec{A} + \vec{B} = \vec{C} \tag{2.18}$$

Note that vector addition proceeds by putting the tail of one at the head of the other, and constructing the vector that completes the triangle. To numerically evaluate the sum of two vectors, we determine their components and add them componentwise, and then reconstruct the total vector:

$$C_x = A_x + B_x \tag{2.19}$$
$$C_y = A_y + B_y \tag{2.20}$$
$$C_z = A_z + B_z \tag{2.21}$$

If we are given a vector in terms of its **length** (magnitude) and **orientation** (direction angle(s)) then we must evaluate its cartesian components before we can add them (for example, in 2D):

$$A_x = \left|\vec{A}\right| \cos(\theta_A) \qquad B_x = \left|\vec{B}\right| \cos\theta_B \tag{2.22}$$
$$A_y = \left|\vec{A}\right| \sin(\theta_A) \qquad B_y = \left|\vec{B}\right| \sin\theta_B \tag{2.23}$$

This process is called **decomposing** the vector into its cartesian components.

The **difference** between two vectors is defined by the addition law. Subtraction is just adding the negative of the vector in question, that is, the vector with the **same** magnitude but the **opposite** direction. This is consistent with the notion of adding or subtracting its components. Note well: Although the vectors themselves may depend upon coordinate system, the difference between two vectors (also called the **displacement** if the two vectors are, for example, the postion vectors of some particle evaluated at two different times) does **not**.

When we reconstruct a vector from its components, we are just using the law of vector addition itself, by **scaling** some special vectors called **unit vectors** and then adding them. Unit vectors are (typically perpendicular) vectors that define the essential directions and orientations of a coordinate system and have unit length. Scaling them involves multiplying these unit vectors by a number that represents the magnitude of the vector component. This scaling number has no direction and is called a **scalar**. Note that the product of a vector and a scalar is always a vector:

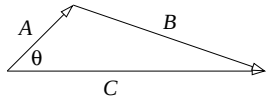
$$\vec{B} = C\vec{A} \tag{2.24}$$

where C is a scalar (number) and $\vec{A}$ is a vector. In this case, $\vec{A}||\vec{B}$.

Finally, we aside from multiplying a scalar and a vector together, we can define products that multiply two vectors together. By "multiply" we mean that if we double the magnitude of either vector, we double the resulting product – the product is *proportional* to the magnitude of either vector. There are two such products for the ordinary vectors we use in this course, and both play *extremely important roles* in physics.

The first product creates a scalar (ordinary number with magnitude but no direction) out of two vectors and is therefore called a **scalar product** or (because of the multiplication symbol chosen) a **dot product**. A scalar is often thought of as being a "length" (magnitude) on a single line. Multiplying two scalars on that line creates a number that has the *units* of length squared but is geometrically not an area. By selecting as a direction for that line the direction of the vector itself, we can use the scalar product to *define* the length of a vector as the *square root* of the vector magnitude times itself:

$$\left|\vec{A}\right| = +\sqrt{\vec{A}\cdot\vec{A}} \tag{2.25}$$



From this usage it is clear that a scalar product of two vectors can never be thought of as an area. If we generalize this idea (preserving the need for our product to be symmetrically proportional to both vectors, we obtain the following definition for the general scalar product:

$$\vec{A} \cdot \vec{B} = A_x * B_x + A_y * B_y \ldots \quad (2.26)$$

$$= \left|\vec{A}\right| \left|\vec{B}\right| \cos(\theta_{AB}) \quad (2.27)$$

This definition can be put into *words* – a scalar product is the length of one vector (either one, say $|\vec{A}|$) times the *component* of the other vector ($|\vec{B}| \cos(\theta_{AB})$ that points in the *same direction* as the vector $\vec{A}$. Alternatively it is the length $|\vec{B}|$ times the component of $\vec{A}$ parallel to $\vec{B}$, $|\vec{A}| \cos(\theta_{AB})$. This product is *symmetric* and *commutative* ($\vec{A}$ and $\vec{B}$ can appear in either order or role).

The other product multiplies two vectors in a way that creates a third vector. It is called a **vector product** or (because of the multiplication symbol chosen) a **cross product**. Because a vector has magnitude and direction, we have to specify the product in such a way that both are defined, which makes the cross product more complicated than the dot product.

As far as magnitude is concerned, we already used the non-areal combination of vectors in the scalar product, so what is left is the product of two vectors that makes an *area* and not just a "scalar length squared". The area of the parallelogram defined by two vectors is just:

$$\text{Area in } \vec{\text{A}} \times \vec{\text{B}} \text{ parallelogram} = \left|\vec{A}\right| \left|\vec{B}\right| sin(\theta_{AB}) \quad (2.28)$$

which we can interpret as "the magnitude of $\vec{A}$ times the component of $\vec{B}$ perpendicular to $\vec{A}$" or vice versa. Let us accept this as the magnitude of the cross product (since it clearly has the proportional property required) and look at the direction.

The area is nonzero only if the two vectors do *not* point along the same line. Since two non-colinear vectors always lie in (or define) a plane (in which the area of the parallelogram itself lies), and since we want the resulting product to be independent of the coordinate system used, one sensible

direction available for the product is along the line *perpendicular to this plane*. This still leaves us with *two* possible directions, though, as the plane has two sides. We have to pick one of the two possibilities by *convention* so that we can communicate with people far away, who might otherwise use a counterclockwise convention to build screws when we used a clockwise convention to order them, whereupon they send us left handed screws for our right handed holes and everybody gets all irritated and everything.

We therefore *define* the direction of the cross product using the *right hand rule*:

> Let the fingers of your *right hand* lie along the direction of the first vector in a cross product (say $\vec{A}$ below). Let them curl naturally through the *small angle* (observe that there are two, one of which is larger than π and one of which is less than π) into the direction of $\vec{B}$. The erect *thumb* of your right hand then points in the general direction of the cross product vector – it at least indicates which of the two perpendicular lines should be used as a direction, unless your thumb and fingers are all double jointed or your bones are missing or you used your left-handed right hand or something.

Putting this all together mathematically, one can show that the following are two equivalent ways to write the cross product of two three dimensional vectors. In components:

$$\vec{A}\times\vec{B} = (A_x*B_y-A_y*B_x)\hat{z}+(A_y*B_z-A_z*B_y)\hat{x}+(A_z*B_x-A_x*B_z)\hat{y} \quad (2.29)$$

where you should note that x, y, z appear in *cyclic order* (xyz, yzx, zxy) in the positive terms and have a minus sign when the order is *anticyclic* (zyx, yxz, xzy). The product is *antisymmetric* and *non-commutative*. In particular

$$\vec{A} \times \vec{B} = -\vec{B} \times \vec{A} \quad (2.30)$$

or the product *changes sign* when the order of the vectors is reversed.

Alternatively in *many* problems it is easier to just use the form:

$$\left|\vec{A} \times \vec{B}\right| = \left|\vec{A}\right| \left|\vec{B}\right| sin(\theta_{AB}) \quad (2.31)$$

to compute the magnitude and assign the direction *literally* by (right) "hand", along the right-handed normal to the AB plane according to the right-hand rule above.

Note that this *axial* property of cross products is realized in nature by things that *twist* or *rotate around an axis*. A screw advances into wood when twisted clockwise, and comes out of wood when twisted counterclockwise. If you let the fingers of your right hand curl around the screw *in the direction of the twist* your *thumb* points in the direction the screw moves, whether it is in or out of the wood. Screws are therefore by convention *right handed*.

One final remark before leaving vector products. We noted above that scalar products and vector products are closely connected to the notions of *length* and *area*, but mathematics per se need not specify the *units* of the quantities multiplied in a product (that is the province of physics, as we shall see). We have numerous examples where two *different* kinds of vectors (with different units but referred to a common coordinate system for direction) are multiplied together with one or the other of these products. In actual fact, there often *is* a buried squared length or area (which we now agree are different kinds of numbers) in those products, but it won't always be obvious in the dimensions of the result.

Two of the most important uses of the scalar and vector product are to define the *work* done as the force through a distance (using a scalar product as work is a scalar quantity) and the *torque* exerted by a force applied at some distance from a center of rotation (using a vector product as torque is an axial vector). These two quantities (work and torque) have the *same units* and yet are very *different* kinds of things. This is just one example of the ways geometry, algebra, and units all get mixed together in physics.

At first this will be very confusing, but remember, back when you where in third grade multiplying *integer numbers* was very confusing and yet rational numbers, irrational numbers, general real numbers, and even complex numbers were all waiting in the wings. This is more of the same, but all of the additions will *mean something* and have a compelling *beauty* that comes out as you study them. Eventually it all makes very, very good sense.

2.5 Functions

One of the most important concepts in algebra is that of the *function*. The formal mathematical definition of the term function [13] is beyond the scope of this short review, but the summary below should be more than enough to work with.

A function is a *mapping* between a set of *coordinates* (which is why we put this section *after* the section on coordinates) and a *single value*. Note well that the "coordinates" in question do *not have to be space and/or time*, they can be any set of *parameters* that are relevant to a problem. In physics, coordinates can be any or all of:

- Spatial coordinates, x, y, z
- Time t
- Momentum p_x, p_y, p_z
- Mass m
- Charge q
- Angular momentum, spin, energy, isospin, flavor, color, and much more, including "spatial" coordinates we *cannot see* in exotica such as string theories or supersymmetric theories.

Note well that many of these things can *equally* well be functions themselves – a potential energy function, for example, will usually return the value of the potential *energy* as a function of some mix of spatial coordinates, mass, charge, and time. Note that the coordinates can be continuous (as most of the ones above are classically) or *discrete* – charge, for example, comes only multiples of e and color can only take on three values.

One formally denotes functions in the notation e.g. $F(\boldsymbol{x})$ where F is the *function name* represented symbolically and $\boldsymbol{x}$ is the entire vector of coordinates of all sorts. In physics we often learn or derive functional forms for important quantities, and may or may not express them as functions in

[13] Wikipedia: http://www.wikipedia.org/wiki/Function (mathematics).

this form. For example, the kinetic energy of a particle can be written either of the two following ways:

$$K(m, \boldsymbol{v}) = \frac{1}{2}mv^2 \tag{2.32}$$

$$K = \frac{1}{2}mv^2 \tag{2.33}$$

These two forms are equivalent in physics, where it is usually "obvious" (at least when a student has studied adequately and accumulated some practical experience solving problems) when we write an expression just what the variable parameters are. Note well that we not infrequently use *non*-variable parameters – in particular constants of nature – in our algebraic expressions in physics as well, so that:

$$U = -\frac{Gm_1m_2}{r} \tag{2.34}$$

is a function of m_1, m_2, and r but includes the gravitational constant $G = 6.67 \times 10^{-11}$ N-m^2/kg^2 in symbolic form. Not all symbols in physics expressions are variable parameters, in other words.

One important property of the mapping required for something to be a true "function" is that there must be only a *single value* of the function for any given set of the coordinates. Two other important definitions are:

Domain The *domain* of a function is the set of all of the coordinates of the function that give rise to unique non-infinite values for the function. That is, for function $f(x)$ it is all of the x's for which f is well defined.

Range The *range* of a function is the set of all values of the function f that arise when its coordinates vary across the entire domain.

For example, for the function $f(x) = \sin(x)$, the domain is the entire real line $x \in (-\infty, \infty)$ and the range is $f \in [-1, 1]$.

Two last ideas that are of great use in solving physics problems algebraically are the notion of *composition* of functions and the *inverse* of a function.

Suppose you are given two functions: one for the potential energy of a mass on a spring:

$$U(x) = \frac{1}{2}kx^2 \tag{2.35}$$

where x is the distance of the mass from its equilibrium position and:

$$x(t) = x_0 \cos(\omega t) \tag{2.36}$$

which is the position as a function of time. We can form the composition of these two functions by substituting the second into the first to obtain:

$$U(t) = \frac{1}{2} k x_0^2 \cos^2(\omega t) \tag{2.37}$$

This sort of "substitution operation" (which we will rarely refer to by name) is an *extremely important* part of solving problems in physics, so keep it in mind at all times!

With the composition operation in mind, we can define the inverse. Not all functions have a unique inverse function, as we shall see, but most of them have an inverse function that we can use *with some restrictions* to solve problems.

Given a function $f(x)$, *if* every value in the range of f corresponds to one and only one value in its domain x, then $f^{-1} = x(f)$ is *also* a function, called the *inverse* of f. When this condition is satisfied, the range of $f(x)$ is the domain of $x(f)$ and vice versa. In terms of composition:

$$x_0 = x(f(x_0)) \tag{2.38}$$

and

$$f_0 = f(x(f_0)) \tag{2.39}$$

for any x_0 in the domain of $f(x)$ and f_0 in the range of $f(x)$ are *both true*; the composition of f and the inverse function for some value f_0 yields f_0 again and is hence an "identity" operation on the range of $f(x)$.

Many functions do not *have* a unique inverse, however. For example, the function:

$$f(x) = \cos(x) \tag{2.40}$$

does not. If we look for values x_m in the domain of this function such that $f(x_m) = 1$, we find an *infinite number*:

$$x_m = 2\pi m \tag{2.41}$$

for $m = 0, \pm 1, \pm 2, \pm 3...$ The mapping is then one value in the range to *many* in the domain and the inverse of $f(x)$ is not a *function* (although we can

still write down an expression for all of the values that each point in the range maps into when inverted).

We can get around this problem by *restricting the domain* to a region where the inverse mapping *is* unique. In this particular case, we can define a function $g(x) = \sin^{-1}(x)$ where the domain of g is only $x \in [-1, 1]$ and the *range* of g is restricted to be $g \in [-\pi/2, \pi/2)$. If this is done, then $x = f(g(x))$ for all $x \in [-1, 1]$ and $x = g(f(x))$ for all $x \in [-\pi/2, \pi/2)$. The inverse function for many of the functions of interest in physics have these sorts of restrictions on the range and domain in order to make the problem well-defined, and in many cases we have some degree of *choice* in the best definition for any given problem, for example, we could use *any* domain of width π that begins or ends on an odd half-integral multiple of π, say $x \in (\pi/2, 3\pi/2]$ or $x \in [9\pi/2, 11\pi/2)$ if it suited the needs of our problem to do so when computing the inverse of $\sin(x)$ (or similar but different ranges for $\cos(x)$ or $\tan(x)$) in physics.

In a related vein, if we examine:

$$f(x) = x^2 \tag{2.42}$$

and try to construct an inverse function we discover two interesting things. First, there are two values in the domain that correspond to each value in the range because:

$$f(x) = f(-x) \tag{2.43}$$

for all x. This causes us to define the inverse function:

$$g(x) = \pm x^{1/2} = \pm\sqrt{x} \tag{2.44}$$

where the sign in this expression selects one of the two possibilities.

The second is that once we have defined the inverse functions for either trig functions or the quadratic function in this way so that they have restricted domains, it is natural to ask: Do these functions have any meaning for the *unrestricted* domain? In other words, if we have defined:

$$g(x) = +\sqrt{x} \tag{2.45}$$

for $x \geq 0$, does $g(x)$ exist for *all* x*?* And if so, what kind of number is g?

This leads us naturally enough into our next section (so keep it in mind) but first we have to deal with several important ideas.

2.5.1 Polynomial Functions

A polynomial function is a sum of monomials:

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n + \ldots \tag{2.46}$$

The numbers $a_0, a_1, \ldots, a_n, \ldots$ are called the *coefficients* of the polynomial.

This sum can be finite and terminate at some n (called the *degree* of the polynomial) or can (for certain series of coefficients with "nice" properties) be infinite and converge to a well defined functional value. Everybody should be familiar with at least the following forms:

$$f(x) = a_0 \quad \text{(0th degree, constant)} \tag{2.47}$$
$$f(x) = a_0 + a_1 x \quad \text{(1st degree, linear)} \tag{2.48}$$
$$f(x) = a_0 + a_1 x + a_2 x^2 \quad \text{(2nd degree, quadratic)} \tag{2.49}$$
$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \quad \text{(3rd degree, cubic)} \tag{2.50}$$

where the first form is clearly independent of x altogether.

Polynomial functions are a simple key to a huge amount of mathematics. For example, differential calculus. It is easy to derive:

$$\frac{dx^n}{dx} = nx^{n-1} \tag{2.51}$$

It is similarly simple to derive

$$\int x^n dx = \frac{1}{n+1} x^{n+1} + \text{constant} \tag{2.52}$$

and we will derive both below to illustrate methodology and help students remember these two *fundamental* rules.

Next we note that many continuous functions can be defined in terms of their *power series* expansion. In fact *any* continuous function can be expanded in the vicinity of a point as a power series, and many of our favorite functions have well known power series that serve as an alternative definition of the function. Although we will not derive it here, one extremely general and powerful way to *compute* this expansion is via the *Taylor series*. Let us define the Taylor series and its close friend and companion, the binomial expansion.

2.5.2 The Taylor Series and Binomial Expansion

Suppose $f(x)$ is a continuous and infinitely differentiable function. Let $x = x_0 + \Delta x$ for some Δx that is "small". Then the following is true:

$$\begin{aligned} f(x_0 + \Delta x) &= f(x)\Big|_{x=x_0} + \frac{df}{dx}\Big|_{x=x_0}\Delta x + \frac{1}{2!}\frac{d^2f}{dx^2}\Big|_{x=x_0}\Delta x^2 \\ &\quad + \frac{1}{3!}\frac{d^3f}{dx^3}\Big|_{x=x_0}\Delta x^3 + \ldots \end{aligned} \tag{2.53}$$

This sum will always converge to the function value (for smooth functions and small enough Δx) if carried out to a high enough degree. Note well that the Taylor series can be rearranged to become the *definition* of the derivative of a function:

$$\frac{df}{dx}\Big|_{x=x_0} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} + \mathcal{O}(\Delta x) \tag{2.54}$$

where the latter symbols stands for "terms of order Δx or smaller" and vanishes in the limit. It can similarly be rearranged to form formal definitions for the second or higher order derivatives of a function, which turns out to be very useful in computational mathematics and physics.

We will find many uses for the Taylor series as we learn physics, because we will *frequently* be interested in the value of a function "near" some known value, or in the limit of very large or very small arguments. Note well that the Taylor series expansion for any polynomial *is* that polynomial, possibly re-expressed around the new "origin" represented by x_0.

To this end we will find it *very* convenient to define the following *binomial expansion*. Suppose we have a function that can be written in the form:

$$f(x) = (c + x)^n \tag{2.55}$$

where n can be *any real or complex number*. We'd like expand this using the Taylor series in terms of a "small" parameter. We therefore factor out the *larger* of x and c from this expression. Suppose it is c. Then:

$$f(x) = (c + x)^n = c^n(1 + \frac{x}{c})^n \tag{2.56}$$

where $x/c < 1$. x/c is now a suitable "small parameter" and we can expand this expression around $x = 0$:

$$f(x) = c^n \left(1 + n\frac{x}{c} + \frac{1}{2!}n(n-1)\left(\frac{x}{c}\right)^2 + \frac{1}{3!}n(n-1)(n-2)\left(\frac{x}{c}\right)^3 + \ldots\right) \quad (2.57)$$

Evaluate the derivatives around $x = 0$ to verify this expansion. Similarly, if x were the larger we could factor out the x and expand in powers of c/x as our small parameter around $c = 0$. In that case we'd get:

$$f(x) = x^n \left(1 + n\frac{c}{x} + \frac{1}{2!}n(n-1)\left(\frac{c}{x}\right)^2 + \frac{1}{3!}n(n-1)(n-2)\left(\frac{c}{x}\right)^3 + \ldots\right) \quad (2.58)$$

Remember, n is *arbitrary* in this expression but you should also verify that if n is any positive integer, the series terminates and you recover $(c+x)^n$ *exactly*. In this case the "small" requirement is no longer necessary.

We summarize both of these forms of the expansion by the part in the brackets. Let $y < 1$ and n be an arbitrary real or complex number (although in this class we will use only n real). Then:

$$(1+y)^n = 1 + ny + \frac{1}{2!}n(n-1)y^2 + \frac{1}{3!}n(n-1)(n-2)y^3 + \ldots \quad (2.59)$$

This is the binomial expansion, and is *very* useful in physics.

2.5.3 Quadratics and Polynomial Roots

As noted above, the purpose of using algebra in physics is so that we can take known expressions that e.g. describe laws of nature and a particular problem and transform these "truths" into a "true" statement of the answer by isolating the *symbol* for that answer on one side of an equation.

For linear problems that is usually either straightforward or impossible. For "simple" linear problems (a single linear equation) it is always possible and usually easy. For sets of simultaneous linear equations in a small number of variables (like the ones represented in the course) one can "always"

use a mix of composition (substitution) and elimination to find the answer desired[14].

What about solving polynomials of higher degree to find values of their variables that represent answers to physics (or other) questions? In general one tries to arrange the polynomial into a *standard form* like the one above, and then finds the *roots* of the polynomial. How easy or difficult this may be depends on many things. In the case of a *quadratic* (second degree polynomial involving at most the square) one can – and we will, below – derive an *algebraic expression* for the roots of an *arbitrary* quadratic.

For third and higher degrees, our ability to solve for the roots is not trivially general. Sometimes we will be able to "see" how to go about it. Other times we won't. There exist computational methodologies that work for most relatively low degree polynomials but for very high degree general polynomials the problem of factorization (finding the roots) is *hard*. We will therefore work through quadratic forms in detail below and then make a couple of observations that will help us factor a few e.g. cubic or quartic polynomials should we encounter ones with one of the "easy" forms.

In physics, quadratic forms are quite common. Motion in one dimension with constant acceleration (for example) quite often requires the solution of a quadratic in time. For the purposes of deriving the quadratic formula, we begin with the "standard form" of a quadratic equation:

$$ax^2 + bx + c = 0 \tag{2.60}$$

(where you should note well that $c = a_0$, $b = a_1$, $c = a_2$ in the general polynomial formula given above).

We wish to find the (two) values of x such that this equation is true, given a, b, c. To do so we must rearrange this equation and *complete the square*.

[14]This is not true in the general case, however. One can, and should, if you are contemplating a physics major, take an *entire college level course* in the methodology of linear algebra in multidimensional systems.

$$\begin{aligned}
ax^2 + bx + c &= 0 \\
ax^2 + bx &= -c \\
x^2 + \frac{b}{a}x &= -\frac{c}{a} \\
x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} &= \frac{b^2}{4a^2} - \frac{c}{a} \\
(x + \frac{b}{2a})^2 &= \frac{b^2}{4a^2} - \frac{c}{a} \\
(x + \frac{b}{2a}) &= \pm\sqrt{\frac{b^2}{4a^2} - \frac{c}{a}} \\
x &= -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a^2} - \frac{c}{a}} \\
x_\pm &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
\end{aligned} \tag{2.61}$$

This last result is the well-known *quadratic formula* and its general solutions are *complex numbers* (because the argument of the square root can easily be negative if $4ac > b^2$). In some cases the complex solution is *desired* as it leads one to e.g. a *complex exponential* solution and hence a trigonometric oscillatory function as we shall see in the next section. In other cases we insist on the solution being real, because if it isn't there is no real solution to the problem posed! Experience solving problems of both types is needed so that a student can learn to recognize both situations and use complex numbers to their advantage.

Before we move on, let us note two cases where we can "easily" solve cubic or quartic polynomials (or higher order polynomials) for their roots algebraically. One is when we take the quadratic formula and multiply it by any power of x, so that it can be *factored*, e.g.

$$\begin{aligned}
ax^3 + bx^2 + cx &= 0 \\
(ax^2 + bx + c)x &= 0
\end{aligned} \tag{2.62}$$

This equation clearly has the two quadratic roots given above plus one (or more, if the power of x is higher) root $x = 0$. In some cases one can factor a solvable term of the form $(x + d)$ by inspection, but this is generally not easy if it is possible at all without solving for the roots some other way first.

The other "tricky" case follows from the observation that:

$$x^2 - a^2 = (x + a)(x - a) \tag{2.63}$$

so that the two roots $x = \pm a$ are solutions. We can generalize this and solve e.g.

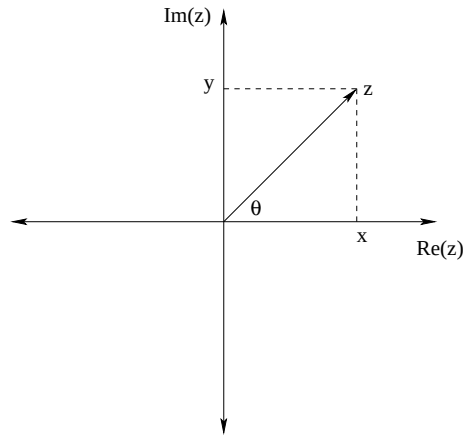
$$x^4 - a^4 = (x^2 - a^2)(x^2 + a^2) = (x - a)(x + a)(x - ia)(x + ia) \tag{2.64}$$

and find the *four* roots $x = \pm a, \pm ia$. One can imagine doing this for still higher powers on occasion.

In *this* course we will almost never have a problem that cannot be solved using "just" the quadratic formula, perhaps augmented by one or the other of these two tricks, although naturally a diligent and motivated student contemplating a math or physics major will prepare for the more difficult future by reviewing the various factorization tricks for "fortunate" integer coefficient polynomials, such as *synthetic division*. However, such a student should *also* be aware that the general problem of finding all the roots of a polynomial of arbitrary degree is *difficult* [15] . So difficult, in fact, that it is known that no *simple* solution involving only arithmetical operations and square roots exists for degree 5 or greater. However it is generally fairly easy to factor arbitrary polynomials to a high degree of accuracy *numerically* using well-known algorithms and a computer.

Now that we understand both inverse functions and Taylor series expansions and quadratics and roots, let us return to the question asked earlier. What happens if we extend the domain of an inverse function outside of the range of the original function? In general we find that the inverse function has no *real* solutions. Or, we can find as noted above when factoring polynomials that like as not there are no real solutions. But that does not mean that solutions do not exist!

[15] Wikipedia: http://www.wikipedia.org/wiki/Polynomial.



2.6 Complex Numbers and Harmonic Trigonometric Functions

We already reviewed very briefly the definition of the unit imaginary number $i = +\sqrt{-1}$. This definition, plus the usual rules for algebra, is enough for us to define both the *imaginary numbers* and a new kind of number called a *complex* number z that is the sum of *real and imaginary parts*, $z = x + iy$.

If we plot the real part of z (x) on the one axis and the imaginary part (y) on another, we note that the complex numbers map into a *plane* that looks *just like* the $x - y$ plane in ordinary plane geometry. Every complex number can be represented as an ordered pair of real numbers, one real and one the magnitude of the imaginary. A picture of this is drawn above.

From this picture and our knowledge of the *definitions* of the trigonometric functions we can quickly and easily deduce some *extremely useful and important* True Facts about:

2.6.1 Complex Numbers

This is a very terse review of their most important properties. From the figure above, we can see that an arbitrary complex number z can *always* be written as:

$$\begin{aligned} z &= x + iy && (2.65)\\ &= |z|\left(\cos(\theta) + i|z|\sin(\theta)\right) && (2.66)\\ &= |z|e^{i\theta} && (2.67) \end{aligned}$$

where $x = |z|\cos(\theta)$, $y = |z|\sin(\theta)$, and $|z| = \sqrt{x^2 + y^2}$. *All complex numbers* can be written as a real amplitude $|z|$ times a complex exponential form involving a phase angle. Again, it is difficult to convey how incredibly useful this result is without further study, but I commend it to your attention.

There are a variety of ways of deriving or justifying the exponential form. Let's examine just one. If we differentiate z with respect to θ we get:

$$\frac{dz}{d\theta} = |z|\left(-\sin(\theta) + i\cos(\theta)\right) = i|z|\left(\cos(\theta) + i\sin(\theta)\right) = iz \qquad (2.68)$$

This gives us a differential equation that is an *identity* of complex numbers. If we multiply both sides by $d\theta$ and divide both sizes by z and integrate, we get:

$$\ln z = i\theta + \text{constant} \qquad (2.69)$$

If we use the inverse function of the natural log (exponentiation of both sides of the equation:

$$\begin{aligned} e^{\ln z} &= e^{(i\theta + \text{constant})} = e^{\text{constant}} e^{i\theta}\\ z &= |z|e^{i\theta} && (2.70) \end{aligned}$$

where $|z|$ is basically a constant of integration that is set to be the *magnitude* of the complex number (or its *modulus*) where the complex exponential piece determines its *complex phase*.

There are a number of really interesting properties that follow from the exponential form. For example, consider multiplying two complex numbers a and b:

$$\begin{aligned} a &= |a|e^{i\theta_a} = |a|\cos(\theta_a) + i|a|\sin(\theta_a) && (2.71)\\ b &= |b|e^{i\theta_b} = |b|\cos(\theta_b) + i|b|\sin(\theta_b) && (2.72)\\ ab &= |a||b|e^{i(\theta_a + \theta_b)} && (2.73) \end{aligned}$$

and we see that multiplying two complex numbers multiplies their *amplitudes* and *adds* their phase angles. Complex multiplication thus *rotates and rescales* numbers in the complex plane.

2.6.2 Trigonometric and Exponential Relations

$$e^{\pm i\theta} = \cos(\theta) \pm i\sin(\theta) \tag{2.74}$$

$$\cos(\theta) = \frac{1}{2}\left(e^{+i\theta} + e^{-i\theta}\right) \tag{2.75}$$

$$\sin(\theta) = \frac{1}{2i}\left(e^{+i\theta} - e^{-i\theta}\right) \tag{2.76}$$

From these relations and the properties of exponential multiplication you can painlessly prove all sorts of trigonometric identities that were immensely painful to prove back in high school

2.6.3 Power Series Expansions

These can easily be evaluated using the Taylor series discussed in the last section, expanded around the origin $z = 0$, and are an alternative way of seeing that $z = e^{i\theta}$. In the case of exponential and trig functions, the expansions converge for all z, not just small ones (although they of course converge *faster* for small ones).

$$e^{x} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \tag{2.77}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots \tag{2.78}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots \tag{2.79}$$

Depending on where you start, these can be used to prove the relations above. They are most useful for getting expansions for small values of their parameters. For small x (to leading order):

$$e^{x} \approx 1 + x \tag{2.80}$$

$$\cos(x) \approx 1 - \frac{x^2}{2!} \tag{2.81}$$

$$\sin(x) \approx x \tag{2.82}$$

$$\tan(x) \approx x \tag{2.83}$$

We will use these fairly often in this course, so learn them.

2.6.4 An Important Relation

A relation I will state without proof that is very important to this course is that the real part of the $x(t)$ derived above:

$$\begin{aligned} \Re(x(t)) &= \Re(x_{0+}e^{+i\omega t} + x_{0-}e^{-i\omega t}) && (2.84) \\ &= X_0\cos(\omega t + \phi) && (2.85) \end{aligned}$$

where ϕ is an arbitrary phase. You can prove this in a few minutes or relaxing, enjoyable algebra from the relations outlined above – remember that x_{0+} and x_{0-} are arbitrary *complex* numbers and so can be written in complex exponential form!

2.7 Calculus

In this section we present a lightning fast review of calculus. It is *most* of what you need to do well in this course.

2.7.1 Differential Calculus

The slope of a line is defined to be the rise divided by the run. For a *curved* line, however, the slope has to be defined *at a point*. Lines (curved or straight, but not infinitely steep) can always be thought of as *functions* of a single variable. We call the slope of a line evaluated at any given point its *derivative*, and call the process of finding that slope *taking the derivative of the function*.

Later we'll say a few words about multivariate (vector) differential calculus, but that is mostly beyond the scope of this course.

The definition of the derivative of a function is:

$$\frac{df}{dx} = \lim_{\Delta x\to 0}\frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{2.86}$$

This is the *slope* of the function at the point x.

First, note that:

$$\frac{d(af)}{dx} = a\frac{df}{dx} \tag{2.87}$$

for any constant a. The constant simply factors out of the definition above.

Second, differentiation is *linear*. That is:

$$\frac{d}{dx}(f(x)+g(x)) = \frac{df(x)}{dx} + \frac{dg(x)}{dx} \tag{2.88}$$

Third, suppose that $f = gh$ (the product of two functions). Then

$$\begin{aligned}
\frac{df}{dx} &= \frac{d(gh)}{dx} = \lim_{\Delta x\to 0} \frac{g(x+\Delta x)h(x+\Delta x) - g(x)h(x)}{\Delta x} \\
&= \lim_{\Delta x\to 0} \frac{\left(g(x)+\frac{dg}{dx}\Delta x)(h(x)+\frac{dh}{dx}\Delta x) - g(x)h(x)\right)}{\Delta x} \\
&= \lim_{\Delta x\to 0} \frac{\left(g(x)\frac{dh}{dx}\Delta x + \frac{dg}{dx}h(x)\Delta x + \frac{dg}{dx}\frac{dh}{dx}(\Delta x)^2)\right)}{\Delta x} \\
&= g(x)\frac{dh}{dx} + \frac{dg}{dx}h(x)
\end{aligned} \tag{2.89}$$

where we used the definition above twice and multiplied everything out. If we multiply this rule by dx we obtain the following rule for the differential of a product:

$$d(gh) = g\ dh + h\ dg \tag{2.90}$$

This is a *very important result* and leads us shortly to integration by parts and later in physics to things like Green's theorem in vector calculus.

We can easily and directly compute the derivative of a mononomial:

$$\begin{aligned}
\frac{dx^n}{dx} &= \lim_{\Delta x\to 0} \frac{x^n + nx^{n-1}\Delta x + n(n-1)x^{n-2}(\Delta x)^2 \ldots + (\Delta x)^n) - x^2}{\Delta x} \\
&= \lim_{\Delta x\to 0} \left(nx^{n-1} + n(n-1)x^{n-2}(\Delta x) \ldots + (\Delta x)^{n-1}\right) \\
&= nx^{n-1}
\end{aligned} \tag{2.91}$$

or we can derive this result by noting that $\frac{dx}{dx} = 1$, the product rule above, and using induction. If one assumes $\frac{dx^n}{dx} = nx^{n-1}$, then

$$\begin{aligned}
\frac{dx^{n+1}}{dx} &= \frac{d(x^n \cdot x)}{dx} \\
&= nx^{n-1}\cdot x + x^n \cdot 1 \\
&= nx^n + x^n = (n+1)x^n
\end{aligned} \tag{2.92}$$

and we're done.

Again it is beyond the scope of this short review to *completely* rederive all of the results of a calculus class, but from what has been presented already one can see how one can systematically proceed. We conclude, therefore, with a simple table of useful derivatives and results in summary (including those above):

$$\frac{da}{dx} = 0 \qquad a \text{ constant} \tag{2.93}$$

$$\frac{d(af(x)}{dx} = a\frac{df(x)}{dx} \qquad a \text{ constant} \tag{2.94}$$

$$\frac{dx^n}{dx} = nx^{n-1} \tag{2.95}$$

$$\frac{d}{dx}(f(x)+g(x)) = \frac{df(x)}{dx} + \frac{dg(x)}{dx} \tag{2.96}$$

$$\frac{df}{dx} = \frac{df}{du}\frac{du}{dx} \qquad \text{chain rule} \tag{2.97}$$

$$\frac{d(gh)}{dx} = g\frac{dh}{dx} + \frac{dg}{dx}h \qquad \text{product rule} \tag{2.98}$$

$$\frac{d(g/h)}{dx} = \frac{\frac{dg}{dx}h - g\frac{dh}{dx}}{h^2} \tag{2.99}$$

$$\frac{de^x}{dx} = e^x \tag{2.100}$$

$$\frac{de^{(ax)}}{dx} = ae^x \qquad \text{from chain rule, } u = ax \tag{2.101}$$

$$\frac{d\sin(ax)}{dx} = a\cos(x) \tag{2.102}$$

$$\frac{d\cos(ax)}{dx} = -a\sin(x) \tag{2.103}$$

$$\frac{d\tan(ax)}{dx} = \frac{a}{\cos^2(ax)} = a\sec^2(ax) \tag{2.104}$$

$$\frac{d\cot(ax)}{dx} = -\frac{a}{\sin^2(ax)} = -a\csc^2(ax) \tag{2.105}$$

$$\frac{d\ln(x)}{dx} = \frac{1}{x} \tag{2.106}$$

$$\tag{2.107}$$

There are a few more integration rules that can be useful in this course, but nearly all of them can be derived in place using these rules, especially the chain rule and product rule.

2.7.2 Integral Calculus

With differentiation under our belt, we need only a few definitions and we'll get integral calculus for free. That's because integration is *antidifferentiation*, the *inverse process* to differentiation. As we'll see, the derivative of a function is unique but its integral has *one free choice* that must be made. We'll also see that the (definite) integral of a function in one dimension is the *area underneath the curve*.

There are lots of ways to facilitate derivations of integral calculus. Most calculus books begin (appropriately) by drawing pictures of curves and showing that the area beneath them can be evaluated by summing small discrete sections and that by means of a limiting process that area is equivalent to the integral of the functional curve. That is, if $f(x)$ is some curve and we wish to find the area beneath a segment of it (from $x = x_1$ to $x = x_2$ for example), one small piece of that area can be written:

$$\Delta A = f(x)\Delta x \tag{2.108}$$

The total area can then be approximately evaluated by piecewise summing N rectangular strips of width $\Delta x = (x_2 - x_1)/N$:

$$A \approx \sum_{n=1}^{N} f(x_1 + n \cdot \Delta x)\Delta x \tag{2.109}$$

(Note that one can get slightly different results if one centers the rectangles or begins them on the low side, but we don't care.)

In the limit that $N \to \infty$ and $\Delta x \to 0$, two things happen. First we note that:

$$f(x) = \frac{dA}{dx} \tag{2.110}$$

by the definition of derivative from the previous section. The function $f(x)$ is the formal derivative of the function representing the area beneath it (independent of the limits as long as x is in the domain of the function.) The second is that we'll get tired adding teensy-weensy rectangles in infinite numbers. We therefore make up a *special symbol* for this infinite limit sum. Σ clearly stands for sum, so we change to another stylized "ess", $\int$, to *also* stand for sum, but now a continuous and infinite sum of all the infinitesimal pieces of area within the range. We now write:

$$A = \int_{x_1}^{x_2} f(x)dx \tag{2.111}$$

as an *exact* result in this limit.

The beauty of this simple approach is that we now can do the following algebra, over and over again, to formulate integrals (sums) of some quantity.

$$\begin{aligned} \frac{dA}{dx} &= f(x) \\ dA &= f(x)dx \\ \int dA &= \int f(x)dx \\ A &= \int_{x_1}^{x_2} f(x)dx \end{aligned} \tag{2.112}$$

This areal integral is called a *definite integral* because it has definite upper and lower bounds. However, we can *also* do the integral with a *variable* upper bound:

$$A(x) = \int_{x_0}^{x} f(x')dx' \tag{2.113}$$

where we indicate how A *varies* as we change x, its upper bound.

We now make a clever observation. $f(x)$ is clearly the function that we get by differentiating this integrated area with a fixed lower bound (which is still arbitrary) with respect to the variable in its upper bound. That is

$$f(x) = \frac{dA(x)}{dx} \tag{2.114}$$

This slope must be the *same* for all possible values of x_0 or this relation would not be correct and unique! We therefore conclude that all the various functions $A(x)$ that can stand for the area differ *only by a constant* (called the constant of integration):

$$A'(x) = A(x) + C \tag{2.115}$$

so that

$$f(x) = \frac{dA'(x)}{dx} = \frac{dA(x}{dx} + \frac{dC}{dx} = \frac{dA(x)}{dx} \tag{2.116}$$

From this we can conclude that the *indefinite* integral of $f(x)$ can be written:

$$A(x) = \int^{x} f(x)dx + A_0 \tag{2.117}$$

where A_0 is the constant of integration. In physics problems the constant of integration must usually be evaluated algebraically from information given in the problem, such as initial conditions.

From this simple definition, we can transform our *existing* table of derivatives into a table of (indefinite) integrals. Let us compute the integral of x^n as an example. We wish to find:

$$g(x) = \int x^n dx \tag{2.118}$$

where we will ignore the constant of integration as being irrelevant to this process (we can and should always add it to one side or the other of any formal indefinite integral unless we can see that it is zero). If we differentiate both sides, the differential and integral are inverse operations and we know:

$$\frac{dg(x)}{dx} = x^n \tag{2.119}$$

Looking on our table of derivatives, we see that:

$$\frac{dx^{n+1}}{dx} = (n+1)x^n \tag{2.120}$$

or

$$\frac{dg(x)}{dx} = x^n = \frac{1}{n+1}\frac{dx^{n+1}}{dx} \tag{2.121}$$

and hence:

$$g(x) = \int^x x^n dx = \frac{1}{n+1}x^{n+1} \tag{2.122}$$

by inspection.

Similarly we can match up the other rules with integral equivalents.

$$\frac{d(af(x))}{dx} = a\frac{df(x)}{dx} \tag{2.123}$$

leads to:

$$\int af(x)dx = a\int f(x)dx \tag{2.124}$$

A very important rule follows from the rule for differentiating a product. If we integrate both sides this becomes:

$$\int d(gh) = gh = \int gdh + \int hdg \tag{2.125}$$

which we often rearrange as:

$$\int gdh = \int d(gh) - \int hdg = gh - \int hdg \tag{2.126}$$

the rule for *integration by parts* which permits us to throw a derivative from one term to another in an integral we are trying to do. This turns out to be very, very useful in evaluating many otherwise extremely difficult integrals.

If we assemble the complete list of (indefinite) integrals that correspond to our list of derivatives, we get something like:

$$\int 0\ dx = 0 + c = c \qquad \text{with } c \text{ constant} \tag{2.127}$$

$$\int af(x)dx = a\int f(x)dx \tag{2.128}$$

$$\int x^n dx = \frac{1}{n+1}x^{n+1} + c \tag{2.129}$$

$$\int (f+g)dx = \int f\ dx + \int g\ dx \tag{2.130}$$

$$\int f(x)dx = \int f(u)\frac{dx}{du}du \qquad \text{change variables} \tag{2.131}$$

$$\int d(gh) = gh = \int gdh + \int hdg \qquad \text{or} \tag{2.132}$$

$$\int gdh = gh - \int hdg \qquad \text{integration by parts} \tag{2.133}$$

$$\int e^x dx = e^x + a \qquad \text{or change variables to} \tag{2.134}$$

$$\int e^{ax} dx = \frac{1}{a}\int e^{ax} d(ax) = \frac{1}{a}e^{ax} + c \tag{2.135}$$

$$\int \sin(ax)dx = \frac{1}{a}\int \sin(ax)d(ax) = \frac{1}{a}\cos(ax) + c \tag{2.136}$$

$$\int \cos(ax)dx = \frac{1}{a}\int \cos(ax)d(ax) = -\frac{1}{a}\sin(ax) + c \tag{2.137}$$

$$\int \frac{dx}{x} = \ln(x) + c \tag{2.138}$$

$$\tag{2.139}$$

It's worth doing a couple of examples to show how to do integrals using these rules. One integral that appears in many physics problems in E&M is:

$$\int_0^R \frac{r\ dr}{(z^2 + r^2)^{3/2}} \tag{2.140}$$

This integral is done using *u substitution* – the chain rule used backwards. We look at it for a second or two and note that if we let

$$u = (z^2 + r^2) \tag{2.141}$$

then

$$du = 2rdr \tag{2.142}$$

and we can rewrite this integral as:

$$\begin{aligned}\int_0^R \frac{r\ dr}{(z^2+r^2)^{3/2}} &= \frac{1}{2}\int_0^R \frac{2r\ dr}{(z^2+r^2)^{3/2}} \\ &= \frac{1}{2}\int_{z^2}^{(z^2+R^2)} u^{-3/2}\ du \\ &= -u^{-1/2}\Big|_{z^2}^{(z^2+R^2)} \\ &= \frac{1}{z} - \frac{1}{(z^2+R^2)^{1/2}}\end{aligned} \tag{2.143}$$

The lesson is that we can often do complicated looking integrals by making a suitable u-substitution that reduces them to a simple integral we know off of our table.

The next one illustrates both integration by parts and doing integrals with infinite upper bounds. Let us evaluate:

$$\int_0^\infty x^2 e^{-ax} dx \tag{2.144}$$

Here we identify two pieces. Let:

$$h(x) = x^2 \tag{2.145}$$

and

$$d(g(x)) = e^{-ax}dx = -\frac{1}{a}e^{-ax}d(-ax) = -\frac{1}{a}d(e^{-ax}) \tag{2.146}$$

or $g(x) = -(1/a)e^{-ax}$. Then our rule for integration by parts becomes:

$$\begin{aligned}\int_0^\infty x^2 e^{-ax} dx &= \int_0^\infty h(x)dg \\ &= h(x)g(x)\Big|_0^\infty - \int_0^\infty g(x)dh \\ &= -\frac{1}{a}x^2 e^{-ax}\Big|_0^\infty + \frac{1}{a}\int_0^\infty e^{-ax}2xdx \\ &= \frac{2}{a}\int_0^\infty xe^{-ax}dx\end{aligned} \tag{2.147}$$

We repeat this process with $h(x) = x$ and with $g(x)$ unchanged:

$$\begin{aligned}\int_0^\infty x^2 e^{-ax} dx &= \frac{2}{a}\int_0^\infty x e^{-ax} dx \\ &= -\frac{2}{a^2} x e^{-ax}\Big|_0^\infty + \frac{2}{a^2}\int_0^\infty e^{-ax} dx \\ &= \frac{2}{a^2}\int_0^\infty e^{-ax} dx \\ &= -\frac{2}{a^3}\int_0^\infty e^{-ax} d(-ax) \\ &= -\frac{2}{a^3} e^{-ax}\Big|_0^\infty = \frac{2}{a^3} \end{aligned} \tag{2.148}$$

If we work a little more generally, we can show that:

$$\int_0^\infty x^n e^{-ax} dx = \frac{(n+1)!}{a^n} \tag{2.149}$$

This is just one illustration of the power of integration by parts to help us do integrals that on the surface appear to be quite difficult.

2.7.3 Vector Calculus

This book will not use a great deal of vector or multivariate calculus, but a *little* general familiarity with it will greatly help the student with e.g. multiple integrals or the idea of the force being the negative gradient of the potential energy. We will content ourselves with a few definitions and examples.

The first definition is that of the *partial derivative*. Given a function of many variables $f(x, y, z...)$, the partial derivative of the function with respect to (say) x is written:

$$\frac{\partial f}{\partial x} \tag{2.150}$$

and is just the regular derivative of the *variable form* of f as a function of all its coordinates with respect to the x coordinate only, holding all the *other* variables constant even if they are not independent and vary in some known way with respect to x.

In many problems, the variables *are* independent and the partial derivative is equal to the regular derivative:

$$\frac{df}{dx} = \frac{\partial f}{\partial x} \tag{2.151}$$

In other problems, the variable y might *depend* on the variable x. So might z. In that case we can form the *total* derivative of f with respect to x by including the variation of f caused by the variation of the other variables as well (basically using the chain rule and composition):

$$\frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} + ... \tag{2.152}$$

Note the different *full* derivative symbol on the left. This is called the "total derivative" with respect to x. Note also that the independent form follows from this second form because $\frac{\partial y}{\partial x} = 0$ and so on are the *algebraic* way of saying that the coordinates are independent.

There are several ways to form vector derivatives of functions, especially *vector* functions. We begin by defining the *gradient* operator, the basic vector differential form:

$$\boldsymbol{\nabla} = \frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}} \tag{2.153}$$

This operator can be applied to a scalar multivariate function f to form its gradient:

$$\boldsymbol{\nabla} f = \frac{\partial f}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial f}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial f}{\partial z}\hat{\boldsymbol{z}} \tag{2.154}$$

The gradient of a function has a magnitude equal to its *maximum* slope at the point in any possible direction, pointing in the direction in which that slope is maximal. It is the "uphill slope" of a curved surface, basically – the word "gradient" *means* slope. In physics this directed slope is *very* useful.

If we wish to take the vector derivative of a vector function there are two common ways to go about it. Suppose $\boldsymbol{E}$ is a vector function of the spatial coordinates. We can form its *divergence*:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \tag{2.155}$$

or its *curl*:

$$\boldsymbol{\nabla} \times \boldsymbol{E} = (\frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y})\hat{\boldsymbol{x}} + (\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z})\hat{\boldsymbol{y}} + (\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x})\hat{\boldsymbol{z}} \tag{2.156}$$

These operations are extremely important in physics courses, especially the more advanced study of electromagnetics, where they are part of the differential formulation of Maxwell's equations, but we will not use them in

a required way in this course. We'll introduce and discuss them and work a rare problem or two, just enough to get the *flavor* of what they mean onboard to front-load a more detailed study later (for majors and possibly engineers or other advanced students only).

2.7.4 Multiple Integrals

The last bit of multivariate calculus we need to address is integration over *multiple* dimensions. We will have many occasions in this text to integrate over *lines*, over *surfaces*, and over *volumes* of space in order to obtain quantities. The integrals themselves are not difficult – in this course they can *always* be done as a series of one, two or three ordinary, independent integrals over each coordinate one at a time with the others held "fixed". This is not always possible and multiple integration can get much more difficult, but we *deliberately* choose problems that illustrate the general idea of integrating over a volume while still remaining accessible to a student with fairly modest calculus skills, no more than is required and reviewed in the sections above.

[Note: This section is not yet finished, but there are examples of all of these in context in the relevant sections below. Check back for later revisions of the book PDF (possibly after contacting the author) if you would like this section to be filled in urgently.]

Part II

Electrostatics

Chapter 3

Introduction

The previous two parts, as one can easily see, are not actual physics. The first is a *very important* review of *how to efficiently learn* physics (or anything else), intended for you (the student) to *read like a novel, one time*, at the very beginning of the course and then refer back to as needed.

The second is a general review of pretty much all of the actual mathematics required for the course – I *assume* that if you're taking physics in college *or* high school you've have had algebra, geometry, trigonometry, and differential and integral calculus (all *required prerequisites* for any sensible calculus-based physics course) but (being no fool) I *also* assume that most students, including ones that got A's in those classes and/or high scores on the relevant advanced placement tests have *forgotten* a lot of it, at least to the point where they are very shaky when it comes to knowing what things mean, how to derive them, or how to apply them to even quite simple problems where the variables all *means* something. If nothing else, putting all the math you might need here in one impossible-to-miss place means you don't have to keep four or five math books handy to get through the problems if you *do* forget (or never learned) something that turns out to be important.

We now depart this "general review" layout (which is obviously not intended to be lectured on, although I usually review the content of the *Preliminaries* on the first day of class at the same time I review the syllabus and set down the class rules and grading scheme that I will use) and embark upon the actual week by week, day by day progress through the course material. For maximal ease of use for you the student and (one hopes) your

instructor, the course is designed to cover *one chapter per week-equivalent*, whether or not the chapter is broken up into a day and a half of lecture (summer school), an hour a day (MWF), or an hour and a half a day (TTh) in a semester based scheme. To emphasize this preferred rhythm, each chapter will be referred to by the *week* it would normally be covered in my own semester-long course.

A week's work in all cases covers just about exactly one "topic" in the course. A very few are spread out over two weeks; one or two compress two related topics into one week, but in all cases the *homework* is assigned on a weekly rhythm to give you ample opportunity to use the *method of three passes* described in the first part of the book, culminating in an expected 2-3 hour *recitation* where you should go over the assigned homework *in a group* of three to six students, with a mentor handy to help you where you get stuck, with a goal of *getting all of the homework perfectly correct by the end of recitation.* That is, at the end of a week plus its recitation, you *should* be able to do *all* of the week's homework, *perfectly*, and *without looking.* You will usually *need* all three passes, the last one working in a group, *plus* the mentored recitation to achieve this degree of competence!

However, *if* you do this, you are almost certain to do well on a quiz that terminates the recitation period, and you will be very likely to *retain* the material and not have to "cram" it in again for the hour exams and/or final exam later in the course. Once you achieve *understanding* and reinforce it with a fair bit of repetition and practice, most students will naturally transform this experience into remarkably deep and permanent learning.

Note well that each week is organized for *maximal ease of learning* with the week/chapter review *first.* Try to *always look at this review before lecture* even if you skip reading the chapter itself until later, when you start your homework. Skimming the whole thing before lecture is, of course, better still. It is a "first pass" that can often make lecture much easier to follow and help free you from the tyranny of note-taking as you only need to note *differences* in the presentation from this text and perhaps the answers to *questions* that helped you understand something during the discussion. Then read or skim it again right before each homework pass.

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*:

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

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

- **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:

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

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

- **The Electrostatic Constant k_e**

  The electrostatic constant k_e sets the scale; it is a *very important number* (as we shall see) – a genuine constant of nature as was G for the gravitational field. It is often expressed in terms of a related quantity called the *permittivity of free space*, ϵ_0, which is more useful for advanced treatments of electrodynamics. We will often/generally use k_e instead in this course (because it is very easy to remember), but

I would like you to know the relationship between this quantity and ϵ_0 so that you can easily calculate the latter if you should ever need it or care.

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

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

- **Electrostatic Field**

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

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

  or

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

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

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

- **Superposition Principle**

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

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

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

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

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

- **Electric Dipoles**

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

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

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

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

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

  Associated with this torque is the following potential energy:

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

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

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

which is actually nontrivial to compute.

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

1.1 Charge

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

Given its ubiquity, it should come as no surprise that the directed observation and study of electricity is quite ancient. It was studied, and written about, at least 3000 years ago, and artifacts that may have been primitive electrical batteries have been discovered in the Middle East that date back to perhaps 250 BCE. 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* [1] 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.

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

[1] Wikipedia: http://www.wikipedia.org/wiki/electric charge.

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

A quantity that is a constant througout all known interactions, neither created nor destroyed, is said (in physics) to be "conserved". In the first semester of this course, you learned of a number of quantities that were *conditionally* conserved – momentum or angular momentum, conserved when the net force or torque acting on a system is zero – or *unconditionally* conserved, such as net energy (or more properly, mass-energy). *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*[2]. 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.

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

[2]Later in the study of physics you may learn of interactions that lead to e.g. *pair production* (or anihillation) – 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	~ 3 MeV
Up antiquark	$\bar{u}$	-2/3	~ 3 MeV
Down quark	d	-1/3	~ 6 MeV
Down antiquark	$\bar{d}$	+1/3	~ 6 MeV
Leptons			
Electron	e^-	-1	511 keV
Positron	e^+	+1	511 keV
Electron neutrino	ν_e	0	< 2 eV

Table 1.1: Charge and Mass of First Generation Fermions

the massive building blocks of nearly everything we see, usually grouped into *families*. One family consists of the *quarks* [3] , which carry a charge that is quantized in units of $e/3$, where $e = 1.6 \times 10^{-19}$ C. The other family are called *leptons* [4] which carry a charge that is quantized in units of e itself.

Table 1.1 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 MeV/c^2 (1.67×10^{-27} kg), almost 2000 times larger than that of an electron at 0.511 MeV/c^2 (9.11×10^{-31} kg). Neutrons are just a hair more massive than a proton (939.6 MeV/c^2). Protons and neutrons are bound together by the strong interaction into an atomic nucleus on the order of 10^{-15} meters in diameter. This (positively charged) nucleus strongly attracts negatively charged electrons via the electrostatic force that is the first object of our study, which then arrange themselves around the nucleus to create a structured, electrically neutral object – the *atom*. Finally, atoms in turn are "glued" together

[3] Wikipedia: http://www.wikipedia.org/wiki/quark.
[4] Wikipedia: http://www.wikipedia.org/wiki/lepton. ,

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 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 his 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:

$$\boldsymbol{F}_{12} = k_e q_1 q_2 \frac{(\boldsymbol{r}_1 - \boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|^3} \tag{1.1}$$

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 $k_e = 9 \times 10^9$ N-m^2/C^2 $= \frac{1}{4\pi\epsilon_0}$ 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 units. It may be simple, but this law is very, very powerful.

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 $\boldsymbol{r}$ is that it is the electrostatic force per unit charge on a small test charge q_0 placed at each point in space $\boldsymbol{r}_0$ in the limit that the test charge vanishes:

$$\boldsymbol{E} = \lim_{q_0 \to 0} \frac{F}{q_0} \tag{1.2}$$

or

$$\boldsymbol{E}(\boldsymbol{r}_0) = kq \frac{(\boldsymbol{r}_0 - \boldsymbol{r})}{|\boldsymbol{r}_0 - \boldsymbol{r}|^3} \tag{1.3}$$

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

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{kq}{r^2}\hat{\boldsymbol{r}} \tag{1.4}$$

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 placed on the field by means of the following rule:

$$\boldsymbol{F} = q\boldsymbol{E} \tag{1.5}$$

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 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

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:

$$\boldsymbol{E}(\boldsymbol{r}) = \sum_i \frac{kq_i(\boldsymbol{r} - \boldsymbol{r}_i)}{|\boldsymbol{r} - \boldsymbol{r}_i|^3} \tag{1.6}$$

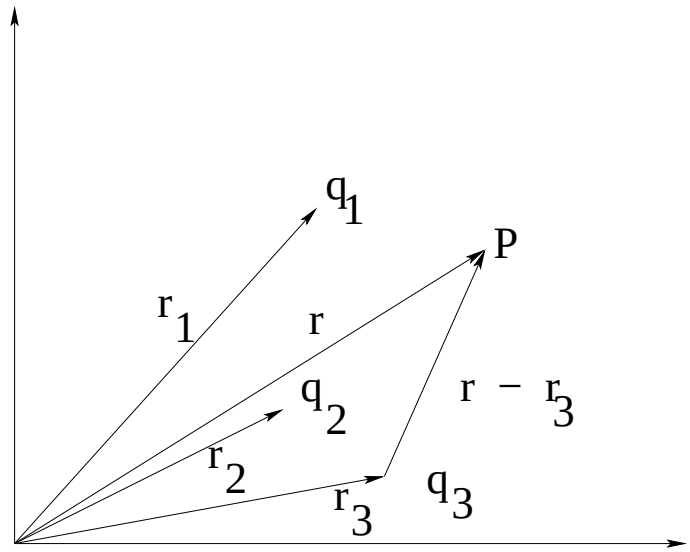


Figure 1.1: Field of Many Charges

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 $\boldsymbol{r}$ inside the material itself very close to one of the consituent 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 an remarkably accurate, accurate as an "average value" even within a material.

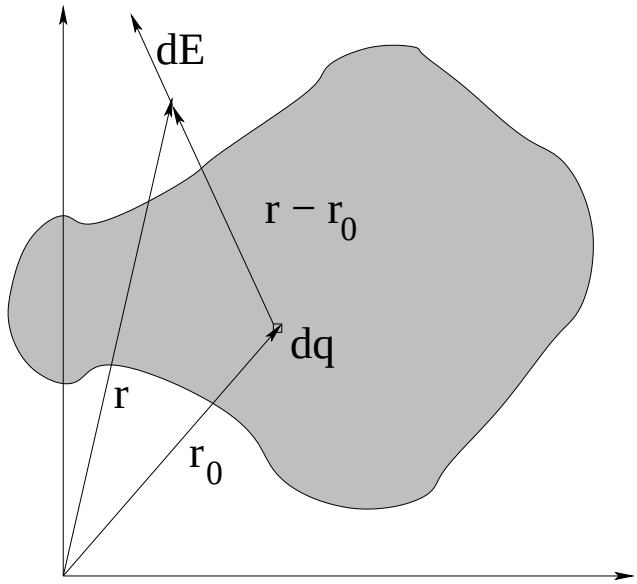


Figure 1.2: Field of Continuous Charge Distribution

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:

$$d\boldsymbol{E}(\boldsymbol{r}) = \frac{k_e\ dq\ (\boldsymbol{r} - \boldsymbol{r}_0)}{|\boldsymbol{r} - \boldsymbol{r}_0|^3} \tag{1.7}$$

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

$$d\boldsymbol{E}(\boldsymbol{r}) = \frac{k_e\ \rho(\boldsymbol{r}_0)\ (\boldsymbol{r} - \boldsymbol{r}_0)d^3r_0}{|\boldsymbol{r} - \boldsymbol{r}_0|^3} \tag{1.8}$$

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

$$\boldsymbol{E}(\boldsymbol{r}) = k_e \int_V \frac{\rho(\boldsymbol{r}_0)(\boldsymbol{r} - \boldsymbol{r}_0)dV_0}{|\boldsymbol{r} - \boldsymbol{r}_0|^3} \tag{1.9}$$

for a 3-dimensional (volume) charge distribution,

$$\boldsymbol{E}(\boldsymbol{r}) = k_e \int_S \frac{\sigma(\boldsymbol{r}_0)(\boldsymbol{r} - \boldsymbol{r}_0)dS_0}{|\boldsymbol{r} - \boldsymbol{r}_0|^3} \tag{1.10}$$

for a surface charge distribution on a surface S, and

$$\boldsymbol{E}(\boldsymbol{r}) = k_e \int_L \frac{\lambda(\boldsymbol{r}_0)(\boldsymbol{r} - \boldsymbol{r}_0)dL_0}{|\boldsymbol{r} - \boldsymbol{r}_0|^3} \tag{1.11}$$

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 $\boldsymbol{r}$ 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.

1.4.1 Example: Field of Two Point Charges

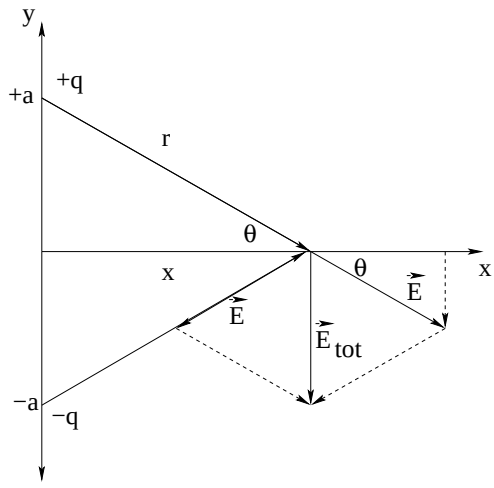


Figure 1.3: Charges $\pm q$ on the y-axis

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:

$$\boldsymbol{E}_{+}(0, y) = \frac{k_e q}{|y - a|}\hat{\boldsymbol{y}} \tag{1.12}$$

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

$$\boldsymbol{E}_{-}(0, y) = -\frac{k_e q}{|y + a|}\hat{\boldsymbol{y}} \tag{1.13}$$

Hence the total field on the y axis is just:

$$\boldsymbol{E}_{\text{tot}}(0, y) = k_e q \left(\frac{1}{|y - a|} - \frac{1}{|y + a|} \right) \hat{\boldsymbol{y}} \tag{1.14}$$

The field on the x-axis is a 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:

$$|\boldsymbol{E}(x,0)| = \frac{k_e q}{r^2} = \frac{k_e q}{(x^2+a^2)} \tag{1.15}$$

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

$$\cos(\theta) = \frac{x}{r} = \frac{x}{(x^2+a^2)^{1/2}} \tag{1.16}$$

$$\sin(\theta) = \frac{a}{r} = \frac{a}{(x^2+a^2)^{1/2}} \tag{1.17}$$

(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:

$$\begin{aligned} E_x &= |\boldsymbol{E}|\cos(\theta) = \frac{k_e q}{(x^2+a^2)} \cdot \frac{x}{(x^2+a^2)^{1/2}} \\ &= \frac{k_e q x}{(x^2+a^2)^{3/2}} \end{aligned} \tag{1.18}$$

and

$$\begin{aligned} E_y &= -|\boldsymbol{E}|\sin(\theta) = -\frac{k_e q}{(x^2+a^2)} \cdot \frac{a}{(x^2+a^2)^{1/2}} \\ &= -\frac{k_e q a}{(x^2+a^2)^{3/2}} \end{aligned} \tag{1.19}$$

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:

$$\boldsymbol{E}_{\rm tot}(x,0) = -2\frac{k_e q a}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}} \tag{1.20}$$

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

$$\boldsymbol{p} = 2qa\hat{\boldsymbol{y}} \tag{1.21}$$

the field can be expressed as:

$$\boldsymbol{E}_{\rm tot}(x,0) = -\frac{k_e|\boldsymbol{p}|}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}} \tag{1.22}$$

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 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:

$$\begin{aligned} \boldsymbol{E}_{\rm tot}(x,0) &= -\frac{k_e|\boldsymbol{p}|}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}} \\ &= -\frac{k_e|\boldsymbol{p}|}{x^3}(1+\left(\frac{a}{x}\right)^2)^{-3/2}\hat{\boldsymbol{y}} \\ &\approx -\frac{k_e|\boldsymbol{p}|}{x^3}\left(1-\frac{3}{2}\left(\frac{a}{x}\right)^2+\ldots\right)\hat{\boldsymbol{y}} \\ &\approx -\frac{k_e|\boldsymbol{p}|}{x^3}\hat{\boldsymbol{y}}+\mathcal{O}\left(\frac{1}{x^5}\right) \end{aligned} \tag{1.23}$$

(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* [5] , 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.

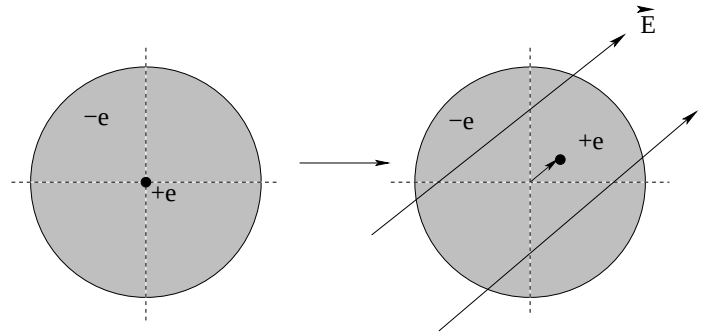


Figure 1.4: Atom with a displaced nucleus forms a dipole

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*. We can model the resulting charge distribution as an electric dipole constructed directly out of two pointlike charges of opposite sign.

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 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

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

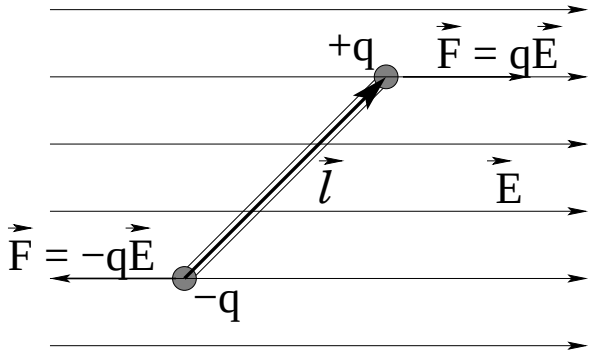


Figure 1.5: A Basic Dipole

two charges is defined to be:

$$\boldsymbol{p} = q\boldsymbol{l} \tag{1.24}$$

where q is the magnitude of the charge and $\boldsymbol{l}$ 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. The electric field of a "point like" dipole has an extremely characteristic shape and a precisely defined functional form in terms of $\boldsymbol{p}$, 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 *torgue* 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:

$$\begin{aligned} \boldsymbol{F} &= -q\boldsymbol{E} + q\boldsymbol{E} \\ &= 0 \end{aligned} \tag{1.25}$$

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:

$$\begin{aligned} \boldsymbol{\tau} &= \boldsymbol{r} \times \boldsymbol{F} \\ &= \boldsymbol{l} \times q\boldsymbol{E} \\ &= q\boldsymbol{l} \times \boldsymbol{E} \\ &= \boldsymbol{p} \times \boldsymbol{E} \end{aligned} \tag{1.26}$$

(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:

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

or

$$U = -\boldsymbol{p} \cdot \boldsymbol{E} \tag{1.27}$$

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 $\boldsymbol{p}$ is a "point dipole", since it assumes that $\boldsymbol{E}$ 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 $\boldsymbol{E}$ varies, so that the value of $\boldsymbol{E}$ "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:

$$\boldsymbol{F} = -\boldsymbol{\nabla} U = \boldsymbol{\nabla}(\boldsymbol{p} \cdot \boldsymbol{E}) \tag{1.28}$$

which can be difficult to compute but is easy to understand. 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 $\boldsymbol{E}$ be the field at (say) the location of the negative charge and $\boldsymbol{E}' = \boldsymbol{E} + \Delta\boldsymbol{E}$ at the location of the positive charge, we have:

$$\begin{aligned} \boldsymbol{F} &= -q\boldsymbol{E} + q\boldsymbol{E}' \\ &= -q\boldsymbol{E} + q\boldsymbol{E} + q\Delta\boldsymbol{E} \\ &= q\Delta\boldsymbol{E} \\ &= \boldsymbol{\nabla}(\boldsymbol{p} \cdot \boldsymbol{E}) \end{aligned} \tag{1.29}$$

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

1.6 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.

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 2.

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. (a) 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. (b) Repeat part (a) for a negative test charge. (c) 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)?

Problem 3.

An electron moves to the right with speed v along the axis of a cathode ray tube. There is an electric field $\boldsymbol{E} = E_0\hat{\boldsymbol{j}}$ in the region between the deflection plates, which are of length l, and everywhere else $\boldsymbol{E} = 0$. 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" into the deflection plates while crossing the deflection zone, and ignore effect of the gravitional 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.

Problem 4.

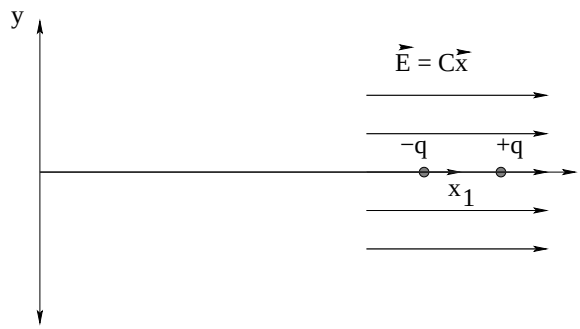


Figure 1.6: Find the force on a diple in a variable field

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 $\boldsymbol{E} = Cx\hat{\boldsymbol{i}}$, where C is a constant. (a) Find the force on the positive charge and that on the negative

charge, and show that the net force on the dipole is $Cp\hat{\boldsymbol{i}}$. (b) Show that in general, if a dipole of moment $\boldsymbol{p}$ lies along the x axis in an electric field in the x direction, the net force on the dipole is given approximately by $\frac{dE_x}{dx}p\hat{\boldsymbol{i}}$.

Problem 5.

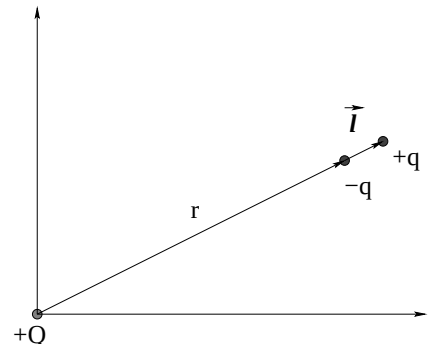


Figure 1.7: Dipole aligned with the field of a point charge.

A positive point charge $+Q$ is at the origin, and a dipole of moment $\boldsymbol{p}$ is at a distance r away and pointing in the radial direction (where $r \gg L$, the physical length of the dipole) as shown. (1) Show that the force exerted on the dipole by the point charge is attractive and has a magnitude $\approx \frac{2kQp}{r^3}$. (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 $\approx \frac{2kp}{r^3}$.

Problem 6.

A ball of known charge q and unknown mass m, initially at rest, falls freely from a height h in a uniform electric field $\boldsymbol{E}$ 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 7.

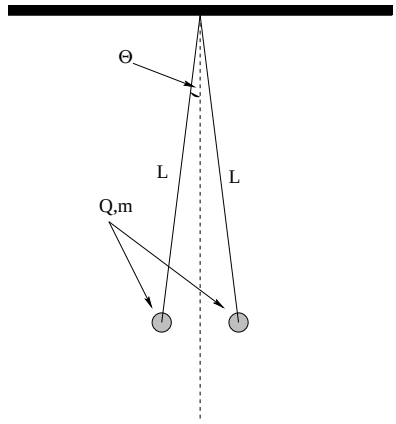


Figure 1.8: An apparatus for verifying Coulomb's Law

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:

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

where k is the electrostatic constant. (b) Find q if $m = 10$ grams, $L = 50$ cm, and $\theta = 10°$. You may use $g = 10$ m/sec^2. 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 8.

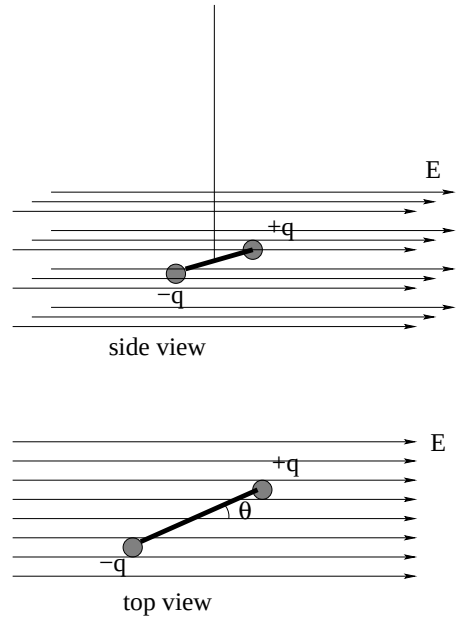


Figure 1.9: Torque on dumbbell-shaped dipole

Suppose you have a dumbbell consisting of two identical 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 $\boldsymbol{E}$. Show that for small values of of the angle θ between the direction of the dipole and the electric field, the system displays simple harmonic motion, and obtain an expression for the period of that motion.

Problem 9.

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. 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 $r = d/2$!.

*** Problem 10.**

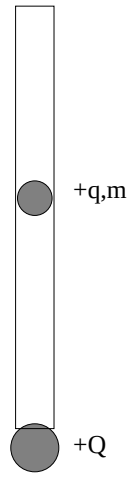


Figure 1.10: Oscillating charge in a vertical tube

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

*** Problem 11.**

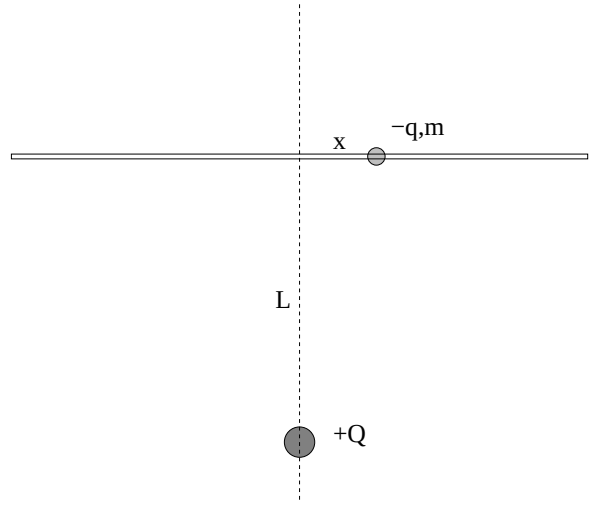


Figure 1.11: Oscillating charge on a frictionless rod

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.

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:

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

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

  Because one has to integrate over the differential *vectors*, this integral is remarkably difficult to perform. We'll revisit it in a much simpler

form when we get to electrostatic *potential*, a scalar quantity that one can usually integrate more easily without this complication.

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

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

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

where in all cases the integral is over the entire charge distribution in question. Note that $dA_0 = d^2 r_0$ 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:

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

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

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

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

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

- 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 $\boldsymbol{E}_\perp$ 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:

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

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

2.1 The Field of Continuous Charge Distributions

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 $\boldsymbol{E}_i$ fields produced by all of these charges. Instead we *average* over the amount of charge inside all of the tiny micron-scale blocks that

might make up a large object. For each block there is a certain *net charge* ΔQ, in the block of size (volume) ΔV. We can use this to define the *average charge density* of the object:

$$\rho = \frac{\Delta Q}{\Delta V} \tag{2.1}$$

Now we can sum over a lot *fewer* objects. There aren't as many blocks a micron in size as there were charges, but there are *still* way, way too many blocks in an object even the size of a centimeter – 10^{12} of them, in fact – too many for us to actually sum up 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:

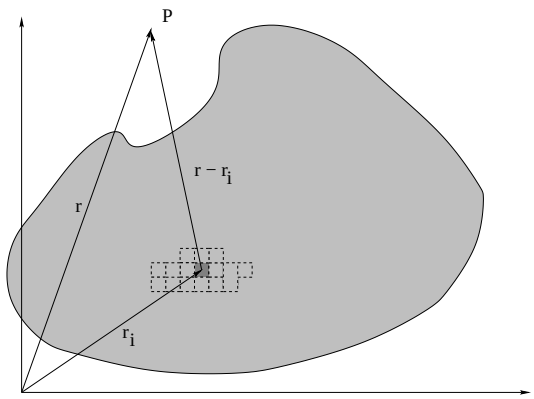


Figure 2.1: Coarse grained average leading to an integral.

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 $\boldsymbol{r}_i$. We can then compute the field using the superposition principle at the point P (position $\boldsymbol{r}$) as:

$$\boldsymbol{E}_{\rm tot}(\boldsymbol{r}) = \sum_i \frac{k\Delta Q_i}{|\boldsymbol{r} - \boldsymbol{r}_i|^2}(\widehat{\boldsymbol{r} - \boldsymbol{r}_i}) \tag{2.2}$$

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:

$$\rho = \lim_{\Delta V \to 0} \frac{\Delta Q}{\Delta V} = \frac{dQ}{dV} \tag{2.3}$$

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):

$$\boldsymbol{E}_{\rm tot}(\boldsymbol{r}) = \sum_i \frac{k\Delta Q_i}{|\boldsymbol{r}-\boldsymbol{r}_i|^2}(\widehat{\boldsymbol{r}-\boldsymbol{r}_i}) = \int_V \frac{k\rho(\boldsymbol{r}')dV'}{|\boldsymbol{r}-\boldsymbol{r}'|^2}(\widehat{\boldsymbol{r}-\boldsymbol{r}'}) \tag{2.4}$$

where we've used $dQ = \rho dV$ (in the primed coordinates we use to replace the $\boldsymbol{r}_i$'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:

$$\boldsymbol{E} = \frac{kq}{r^2}\hat{\boldsymbol{r}} \quad \Longleftrightarrow \quad d\boldsymbol{E} = \frac{k\,dq}{r^2}\hat{\boldsymbol{r}} \tag{2.5}$$

To use the latter, we just have to find dq for the particular kind of distribution, and be able to do the final integrals.

We used charge per unit volume in this discussion, but we will find that charge often distributes itself on surfaces, and we'll often need to find the field produced by lines as well. We therefore define all of the charge densities we might need to handle these cases as:

$$\rho = \frac{dq}{dV} \quad \Longleftrightarrow \quad dq = \rho\, dV \tag{2.6}$$

$$\sigma = \frac{dq}{dA} \quad \Longleftrightarrow \quad dq = \rho\, dA \tag{2.7}$$

$$\lambda = \frac{dq}{d\ell} \quad \Longleftrightarrow \quad dq = \rho\, d\ell \tag{2.8}$$

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:

1. 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 $d\boldsymbol{E}$ as given above.

2. Put down an expression for $d\boldsymbol{E}$ (or rather, usually $|d\boldsymbol{E}|$) in terms of the coordinates, and find its *vector* components in terms of those same coordinates, using symmetry to eliminate unnecessary work.

3. Do the integral(s), find the field $\boldsymbol{E}$ at the desired point.

The first two are pretty simple, and are worth most of the credit. The last will be easy enough if you've done the homework and are working hard to relearn all the calculus you need to do the integrals required in this course, 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.

2.1.1 Example: Circular Loop of Charge

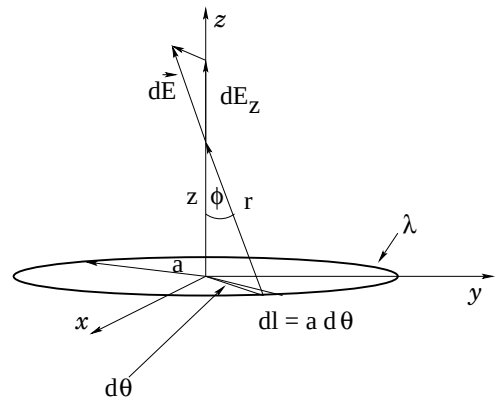


Figure 2.2: A charged ring with charge per unit length λ.

In figure 2.2 above we see a circular ring of charge of radius a and uniform charge per unit length:

$$\lambda = \frac{Q}{L} = \frac{Q}{2\pi a} \tag{2.9}$$

Our job is to find the electric field at an arbitrary point on the z-axis, a point with sufficient symmetry to make the evaluation fairly straightforward[1].

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:

$$dq = \lambda\ d\ell = \lambda a\ d\theta = Q\frac{d\theta}{2\pi} \tag{2.10}$$

where the last form is clearly the *fraction* of the total charge that lies inside the tiny subtended arc. The magnitude of the field produced by this little

[1] We *could* use the same general approach to find the field at an arbitrary point in space, but the *calculus* and *geometry* required to get an actual 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.

chunk of charge at the point z on the axis is:

$$|d\boldsymbol{E}| = \frac{k_e dq}{r^2} = \frac{k_e \lambda a d\theta}{z^2 + a^2} \tag{2.11}$$

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:

$$dE_x = -|d\boldsymbol{E}| \sin\phi \cos\theta \tag{2.12}$$

$$dE_y = -|d\boldsymbol{E}| \sin\phi \sin\theta \tag{2.13}$$

$$dE_z = |d\boldsymbol{E}| \cos\phi \tag{2.14}$$

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

$$\sin\phi = \frac{a}{r} = \frac{a}{(z^2 + a^2)^{1/2}} \tag{2.15}$$

$$\cos\phi = \frac{z}{r} = \frac{z}{(z^2 + a^2)^{1/2}} \tag{2.16}$$

so that:

$$E_z = \int_0^{2\pi} \frac{k_e \lambda z \; a d\theta}{(z^2 + a^2)^{3/2}} = \frac{k_e \; \lambda(2\pi a) \; z}{(z^2 + a^2)^{3/2}} = \frac{k_e Q \; z}{(z^2 + a^2)^{3/2}} \tag{2.17}$$

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

$$E_x = -\int_0^{2\pi} \frac{k_e \lambda a^2 \cos\theta d\theta}{(z^2 + a^2)^{3/2}} = -\frac{k_e \lambda a^2}{(z^2 + a^2)^{3/2}} \cdot \sin\theta \Big|_0^{2\pi} = 0 \tag{2.18}$$

(and ditto, of course, for E_y)!

2.1.2 Example: Long Straight Line of Charge

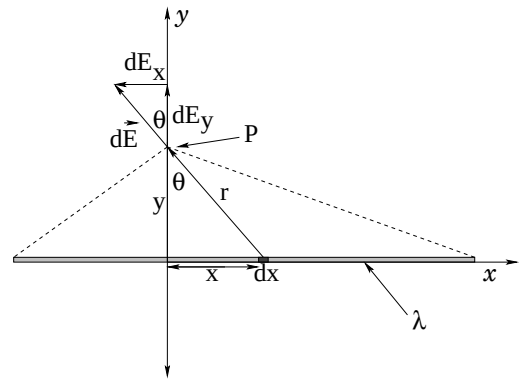


Figure 2.3: A straight line of charge with uniform charge per unit length λ.

In figure 2.3 we see a long straight line of charge. As before, we have to choose a coordinate system in terms of which to do the integral to add up the field components produced by all the little chunks of charge that make up the line.

At first glance, it seems as though cartesian components are a natural choice for the problem, so we start by using them. We want to find the field at an arbitrary point P in space, so we pick one 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:

$$dq = \lambda \, dx \tag{2.19}$$

Finally, the magnitude of the field is given by:

$$|d\boldsymbol{E}| = \frac{k_e dq}{r^2} = \frac{k_e \lambda \, dx}{(x^2 + y^2)} \tag{2.20}$$

We need in this case to evaluate *both* 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:

$$dE_x = -|d\boldsymbol{E}| \sin\theta \tag{2.21}$$

$$dE_y = |d\boldsymbol{E}| \cos\theta \tag{2.22}$$

where we will assume that the θ we have drawn is *positive* when swung out to the right in the positive x direction, and negative when it swings out in the direction of negative x. Noting that $\cos\theta = y/r$ we get:

$$dE_y = \frac{k_e\lambda\ dx}{r^2}\cos\theta = \frac{k_e\lambda\ dx}{(x^2+y^2)}\cos\theta = k_e\lambda y\frac{dx}{(x^2+y^2)^{3/2}} \tag{2.23}$$

(for example). This, unfortunately, doesn't look terribly easy to integrate!

In fact, this is one of the most difficult integrals we have to do in this course, not because it is *particularly* difficult but because it is one of the few times we have to integrate something other than $x^n dx$, a simple trig function, or an exponential function. 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:

$$x = y\tan\theta \tag{2.24}$$

so

$$dx = \frac{y\ d\theta}{\cos^2\theta} \tag{2.25}$$

and

$$r = \frac{y}{\cos\theta} \tag{2.26}$$

If we substitute these into the expressions above we get:

$$dE_y = \frac{k_e\lambda\ dx}{r^2}\cos\theta = k_e\lambda\left(\frac{y\ d\theta}{\cos^2\theta}\right)\left(\frac{\cos^2\theta}{y^2}\right)\cos\theta = \frac{k_e\lambda}{y}\cos\theta d\theta \tag{2.27}$$

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, $theta_2$ on the right. Thus:

$$E_y = \frac{k_e\lambda}{y}\int_{\theta_1}^{\theta_2}\cos\theta d\theta = \frac{k_e\lambda}{y}(\sin\theta_2 - \sin\theta_1) \tag{2.28}$$

(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:

$$E_x = -\frac{k_e\lambda}{y}\int_{\theta_1}^{\theta_2}\sin\theta d\theta = \frac{k_e\lambda}{y}(\cos\theta_2 - \cos\theta_1) \tag{2.29}$$

An interesting consequence of this result is that we can easily evaluate the field a distance y away from an *infinite* line of charge (that still has a uniform charge per unit length λ. In that case, $\theta_1 = -\pi/2$ and $\theta_2 = \pi/2$. We get:

$$E_x(\infty) = 0 \tag{2.30}$$

$$E_y(\infty) = \frac{2k_e\lambda}{y} \tag{2.31}$$

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

2.1.3 Example: Circular Disk of Charge

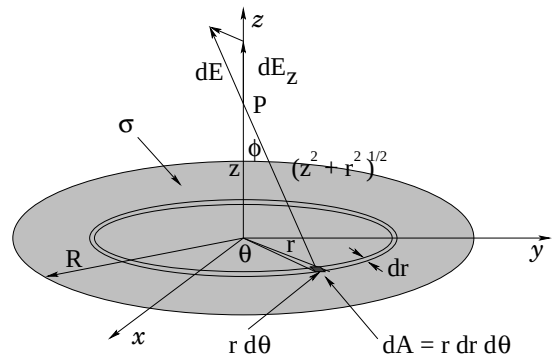


Figure 2.4: A charged disk with charge per unit area σ.

In figure 2.4 above we see a disk of charge with a uniform charge density:

$$\sigma = \frac{Q}{\pi R^2} \tag{2.32}$$

As before with a ring, we can only easily evaluate the field on the z-axis where we 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:

$$A = \int dA = \int_0^R \int_0^{2\pi} r dr d\theta = \left(\int_0^R r dr\right)\left(\int_0^{2\pi} d\theta\right) = \frac{R^2}{2}(2\pi) = \pi R^2 \tag{2.33}$$

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:

$$dq = \sigma dA = \sigma r dr d\theta = \frac{Q}{\pi R^2} r dr d\theta \tag{2.34}$$

As before, we find

$$|d\boldsymbol{E}| = \frac{k_e dq}{(r^2 + z^2)} = \frac{k_e \sigma\ r dr\ d\theta}{(r^2 + z^2)} \tag{2.35}$$

and

$$dE_z = |d\boldsymbol{E}| \cos\phi = \frac{k_e \sigma z\ r dr\ d\theta}{(r^2 + z^2)^{3/2}} \tag{2.36}$$

Finally:

$$E_z = \int dE_z = k_e \sigma z \int_0^R \int_0^{2\pi} \frac{r dr\ d\theta}{(r^2 + z^2)^{3/2}} \tag{2.37}$$

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

$$\begin{aligned} E_z &= 2\pi k_e \sigma z \int_0^R \frac{r dr}{(r^2 + z^2)^{3/2}} \\ &= \pi k_e \sigma z \int_0^R (r^2 + z^2)^{-3/2} (2r dr) \\ &= -2\pi k_e \sigma z (r^2 + z^2)^{-1/2}\Big|_0^R \\ &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2 + z^2)^{1/2}}\right) \\ &= 2\pi k_e \sigma\,(1 - \cos\Phi) \end{aligned} \tag{2.38}$$

where (as was pointed out to me by one of my many clever students) $\cos Phi = z/\sqrt{R^2+z^2}$ where the angle Φ points from P to the edge of the disk.

There are two useful limits for us to explore for this problem. One is the limit that $R \to \infty$ (which we can also interpret as $\Phi \to \pi/2$). In this limit, the disk of charge is *infinite* in extent – it is an infinite plane of uniform charge. The field is obviously:

$$E_z(\infty) = 2\pi k_e \sigma \tag{2.39}$$

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:

$$\begin{aligned}
E_z &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2+z^2)^{1/2}}\right) \\
&= 2\pi k_e \sigma \left(1 - \frac{z}{z(1+\frac{R^2}{z^2})^{1/2}}\right) \\
&= 2\pi k_e \sigma \left(1 - (1+\frac{R^2}{z^2})^{-1/2}\right) \\
&\approx 2\pi k_e \sigma \left(1 - (1 - \frac{1}{2}\frac{R^2}{z^2} + ...)\right) \\
&\approx \pi k_e \sigma \left(\frac{R^2}{z^2}\right) \\
&\approx \frac{k_e(\pi R^2 \sigma)}{z^2} \\
&\approx \frac{k_e Q}{z^2}
\end{aligned} \tag{2.40}$$

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.

2.1.4 Example: Sphere of Charge

At some point I will show how you can find the field of a uniform spherical ball (or spherical shell) of charge by direct integration, but the integrals in this case are *not* easy to set up (although they are easy enough to do once they are set up). One has to choose a spherical polar coordinate system, choose a point P on the z-axis, and do some clever changes of variables to reduce the integral to something that can reasonably easily be done. This is more of interest to physics majors or other students who are seeking to develop mad skills in mathematics than it is to general physics students, so it is pretty safe to skip it here.

Besides, we'll get this result *trivially simply* in the next section on Gauss's Law!

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* [2] . 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

[2]Wikipedia: http://www.wikipedia.org/wiki/Maxwell's Equations.

Coulomb's Law in disguise) and the superposition principle. Derivations, of course, work both ways and physicists today generally consider Gauss's Law the fundamental law of nature and the field of a point charge and Coulomb's law are rather consequences to be derived from it instead of the other way around. You will not be responsible for being able to "do" the derivation yourself in a problem or on an exam, but it is strongly advised that you work through it a couple of times anyway and get to where you intuitively understand the relationship between flux integrals and conservation, as we'll use this idea in a critical way later when we add the Maxwell Displacement Current to Ampere's Law in order to be able to show that light is an electromagnetic wave!

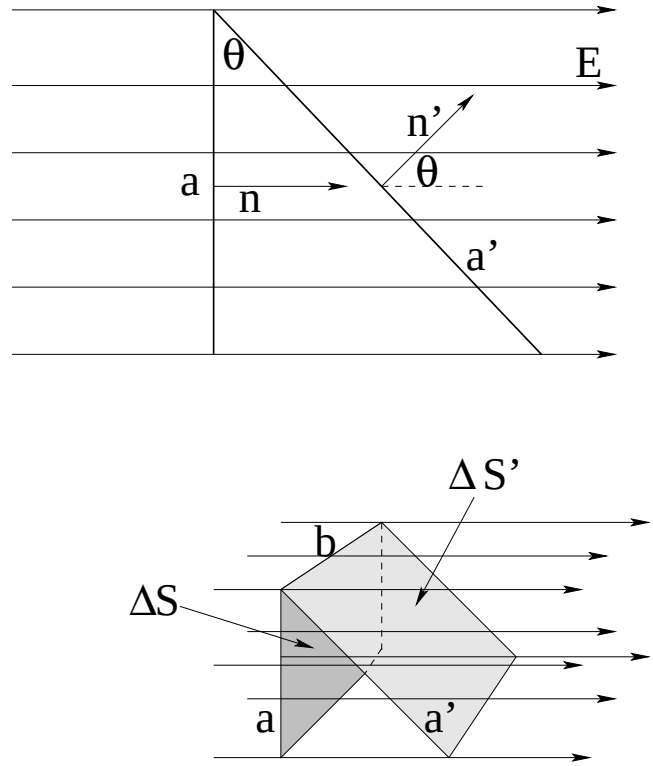


Figure 2.5: Geometry of the flux integral over a small surface area

We begin our derivation of Gauss's Law by considering the *flux* of the electrostatic vector field through a small rectangular patch of surface ΔS. To compute this, we first must understand what the flux of an arbitrary vector field $\boldsymbol{F}$ through a surface S *is*. Mathematically, the flux of a vector

field through some surface is defined to be:

$$\phi_f = \int_{\Delta S} \boldsymbol{F} \cdot \hat{n}\, dS \tag{2.41}$$

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 $\boldsymbol{F}$ at all points on the surface, computes the component of $\boldsymbol{F}$ that goes *through* the surface at right angles (instead of tangent to the surface, which doesn't really go "through"), multiplies it times a tiny differential chunk of the area, and then adds up all the differential chunks thus computed.

Let's look at this in more detail, specializing to the case of the electric field. Consider figure 2.5, where we show electric field lines flowing through a small $\Delta S = ab$ at right angles to the field lines (so that a unit vector $\hat{\boldsymbol{n}}$ normal to the surface is *parallel* to the electric field). ΔS is small enough that the continuous field is approximately uniform across it (we will eventually make it differentially small, of course, so this is no problem).

Since the field is uniform and at right angles to the field, the flux through just this little chunk is easy to evaluate. It is just:

$$\Delta\phi_e = |\boldsymbol{E}|\Delta S = |\boldsymbol{E}|ab \tag{2.42}$$

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 $\boldsymbol{E}$ as shown. Basically, all the field lines that pass through the first surface pass through the second surface, and again we are assuming that the field is continuous and we can make the picture as small as we like (differentially small in the limit) so that a conserved $\boldsymbol{E}$ doesn't change its *magnitude or direction* in between the two surfaces.

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

$$\Delta S' = a'b = \frac{ab}{\cos(\theta)} \tag{2.43}$$

If we just multiply $|\boldsymbol{E}|$ by $\Delta S'$, we see that we'll get $\Delta\phi'_e = \Delta\phi_e/\cos(\theta)$, right? And we'd like to get the same thing, as we'd like the flux integral

to *measure* the continuity and conservation of the electric field across the tiny region between the two surfaces. So we multiply by $\cos(\theta)$ on top to compensate and get:

$$\begin{aligned} \Delta\phi'_e &= |vE|\cos(\theta)a'b \\ &= |vE|\cos(\theta)\frac{ab}{\cos(\theta)} \\ &= |vE|ab \\ &= \Delta\phi_e \end{aligned} \tag{2.44}$$

We can interpret this as meaning (in words) "If $\boldsymbol{E}$ is a continuous, constant vector field in the region between ΔS and $\Delta S'$, then $\Delta\phi'_e = \Delta\phi_e$ and the flux through the two surfaces is conserved."

Note that $|\boldsymbol{E}| = \boldsymbol{E}\cdot\hat{\boldsymbol{n}}$ and $|\boldsymbol{E}|\cos(\theta) = \boldsymbol{E}\cdot\hat{\boldsymbol{n}}'$, so that we can write:

$$\begin{aligned} \lim_{\Delta S\to 0}\Delta\phi_e &= \boldsymbol{E}\cdot\hat{\boldsymbol{n}}\Delta S \\ d\phi_e &= \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dS \end{aligned} \tag{2.45}$$

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 $\boldsymbol{E}$.

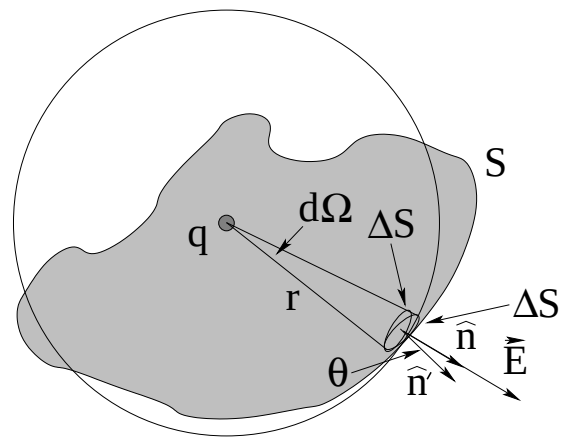


Figure 2.6: Point charge inside a closed surface S. Note that the flux through the tipped differential piece of the surface $\Delta S' = r^2\ d\Omega/\cos\theta$ is equal to that through the *untipped* spherical piece of the surface $\Delta S = r^2\ d\Omega$ that is subtended by the same solid angle $d\Omega$ 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:

$$\boldsymbol{E} = \frac{k_e q \hat{\boldsymbol{r}}}{r^2} \tag{2.46}$$

This situation is pictured in figure 2.6.

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 $\boldsymbol{E}$ will be *exactly* constant and *exactly* perpendicular to that patch. In fact, the flux through that surface patch is:

$$\Delta\phi_e = \boldsymbol{E} \cdot \hat{\boldsymbol{n}} \Delta S = |\boldsymbol{E}| r^2 \Delta\Omega \tag{2.47}$$

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:

$$\Delta\phi'_e = \boldsymbol{E} \cdot \hat{\boldsymbol{n}}' \Delta S' = |\boldsymbol{E}| \cos\theta \frac{r^2 \Delta\Omega}{\cos\theta} = |\boldsymbol{E}| r^2 \Delta\Omega = \Delta\phi_e \tag{2.48}$$

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

$$\begin{aligned} d\phi_e &= \boldsymbol{E} \cdot \hat{\boldsymbol{n}} dS' \\ &= |\boldsymbol{E}| r^2 d\Omega \\ &= \frac{k_e q}{r^2} r^2 d\Omega \\ &= k_e q \; d\Omega \end{aligned} \tag{2.49}$$

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:

$$\int d\Omega = \int_0^{\pi} \int_0^{2\pi} \sin(\theta) d\theta \; d\phi = 4\pi \;\; \text{steradians} \tag{2.50}$$

we get:

$$\phi_e = \oint_{S'} \boldsymbol{E} \cdot \hat{\boldsymbol{n}}\, dS = 4\pi k_e\, q \tag{2.51}$$

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 $\hat{\boldsymbol{n}}$ 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

$$\oint_{S/V} \boldsymbol{E} \cdot \hat{\boldsymbol{n}}\, dA = 4\pi k_e \int_V \rho\, dV = \frac{Q_{\text{in S}}}{\epsilon_0} \tag{2.52}$$

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 planarly symmetric, we can do the flux integral in Gauss's law *once and for all* for the symmetry. As we'll see, all that remains for us to be able to easily obtain the field from algebra is for us to evaluate the total charge inside a Gaussian surface for any given symmetric distribution. Here's the recipe:

1. 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).

2. 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!

3. 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.

4. Set the (once and for all) flux integral equal to the (computed per problem) charge inside the surface and solve for $|\boldsymbol{E}|$. 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.

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 2.7, there are two distinct regions where we must find the field: *inside* the shell and *outside* the shell. Draw a *spherical*

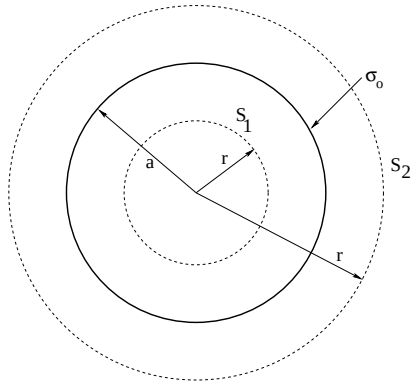


Figure 2.7: A spherical shell of radius a, carrying a uniform charge per unit area σ_0. Two spherical concentric *Gaussian surfaces* S_1 (with radius $r < a$ and S_2 (with radius $r > a$) are shown.

Gaussian surface S_1 inside the sphere (for $r < a$). From the symmetry of the distribution we know that the field $\boldsymbol{E}$ must point in the direction of $\boldsymbol{r}$ and (hence) be perpendicular and constant in magnitude at all points on the Gaussian surface S_1. Hence:

$$\phi_e = \oint_{S_1} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\, dA = E_r \oint_{S_1} dA = E_r(4\pi r^2) \tag{2.53}$$

where it is presumed that everybody knows how to integrate to evalute the area of a sphere *and* knows the result.

The total charge Q_S inside this sphere is *zero* by inspection – the fingers and toes thing. That was easy! Now we write Gauss's law:

$$\phi_e = \oint_{S_1} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\, dA = E_r(4\pi r^2) = \frac{Q_{S_1}}{\epsilon_0} = 0 \tag{2.54}$$

and solve for E_r:

$$\begin{aligned} E_r(4\pi r^2) &= 0 \\ &= \frac{0}{4\pi r^2} \\ E_r &= 0 \qquad \text{for } r < a \end{aligned} \tag{2.55}$$

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*:

$$\phi_e = \oint_{S_2} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\ dA = E_r \oint_{S_2} dA = E_r(4\pi r^2) \tag{2.56}$$

and in fact it will *always* have this algebraic form for a spherical problem, to the point where we will get bored writing this line out umpty times doing homework. Don't let that stop you! Do it every time, as when you know something well enough to be slightly bored writing it out, that's just about perfect, isn't it?

Again we can count up the charge inside S_2 on the thumbs of one hand. It is the total charge on the shell! Which is, in fact (noting that dA for a spherical shell of radius a is $a^2 \sin(\theta)\ d\theta\ d\phi$):

$$\begin{aligned} Q_S &= \int_S \sigma_0\ dA = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta\ a^2\sigma_0 = 2\pi a^2\sigma_0 \int_{-1}^{1} d(\cos\theta) \\ &= 4\pi a^2\sigma_0 \end{aligned} \tag{2.57}$$

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'll have occasion to use again in other problems.

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

$$\phi_e = E_r(4\pi r^2) = \frac{Q_S}{\epsilon_0} \tag{2.58}$$

or

$$E_r = \frac{Q_s}{4\pi\epsilon_0\ r^2} = \frac{k_e Q_s}{r^2} \tag{2.59}$$

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.*

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!

2.3.2 Electric Field of a Solid Sphere of Charge

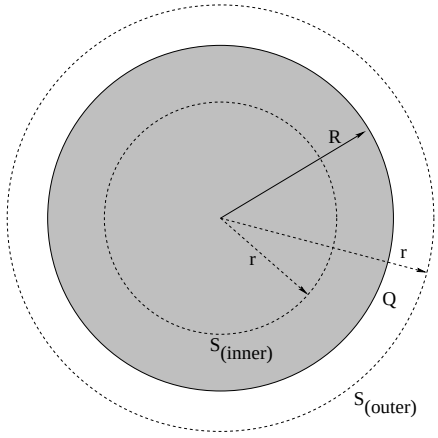


Figure 2.8: A solid sphere of uniform charge density ρ and radius R.

> 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.

We begin by writing Gauss's Law for the outer surface in the figure 2.8:

$$\begin{aligned}
\oint_{S_{\text{outer}}} \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA &= 4\pi k_e \int_{V/S} \rho dV \\
E_r 4\pi r^2 &= 4\pi k_e \left\{ \int_0^R \int_0^{2\pi} \int_0^{\pi} \rho r^2 \sin(\theta)\; d\theta\, d\phi\, dr \right. \\
&\qquad \left. + \int_R^r 0\; dV \right\} \\
&= 4\pi k_e (2\pi\rho) \int_0^R r^2 dr \int_{-1}^{1} d(\cos(\theta)) \\
&= 4\pi k_e (\frac{4\pi R^3}{3}\rho) \\
&= 4\pi k_e Q_{\text{total}} \qquad (2.60)
\end{aligned}$$

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

$$E_r = \frac{k_e Q}{r^2} \qquad r > R \qquad (2.61)$$

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 $\theta \to \cos(\theta)$ (and change and swap order of the limits to get rid of the minus sign). It is *very often* much easier to integrate with $\cos(\theta)$ as the variable instead of θ in spherical

coordinates – in this case one can just look at it and see that one gets "2" from the integral in your head, for example.

Now we redo the whole thing for the interior integral:

$$\begin{aligned}
\oint_{S_{\text{inner}}} \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA &= 4\pi k_e \int_{V/S} \rho dV \\
E_r 4\pi r^2 &= 4\pi k_e \int_0^r \int_0^{2\pi} \int_0^{\pi} \rho r'^2 \sin(\theta)\; d\theta\, d\phi\, dr' \\
&= 4\pi k_e (2\pi\rho) \int_0^r r'^2 dr' \int_{-1}^{1} d(\cos(\theta)) \\
&= 4\pi k_e (\frac{4\pi r^3}{3}\rho)
\end{aligned} \tag{2.62}$$

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

$$E_r = k_e \left(\frac{4\pi\rho r}{3}\right) \qquad r < R \tag{2.63}$$

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

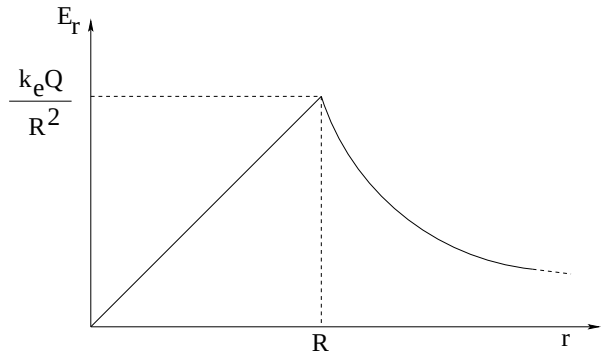


Figure 2.9: Electric field produced by a uniform sphere of charge both inside and outside, as a function of r.

linearly inside the sphere and is *zero* at the origin, not infinite! Outside, the field drops off like $1/r^2$ – as you do more and more of these, you'll come to expect this to the point where you don't think twice about it. Any charge distribution with compact support and a net charge (spherical or not) produces a field that is dominantly *monopolar* and drops off like $1/r^2$ far away from the distribution.

This is very cool! The fact that the field is bounded at the origin means that the *singularity* that appears implicitly in the electrostatic field of a *point* charge need not trouble us if the charge isn't really a *point* charge but is rather a small ball of charge. However, if charge is bound up in a small finite size ball it produces *other* problems – such as the need for a force to hold it all together, as electrostatic charge of a single sign *repels itself.* In the case of a proton, there *is* such a binding force – the strong nuclear force. In the case of electrons, quarks, elementary particles, there *is* (as far as we can tell experimentally or predict theoretically) no such force, and hence those particles "should" be, and experimentally appear to be, truly *pointlike.* Which leads to a whole new set of problems (oops, that nasty infinity is back and has to be dealt with), the invention of renormalizable quantum field theories that soften or throw away the infinity – and in the process, makes physics an *enormously interesting discipline!* Much as we *do* understand at this point, the problem of understanding our Universe, especially at the smallest length and time scales, is far from solved.

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 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 – all that separates you is a bit more work with the damped driven harmonic oscillator. 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.

2.3.3 Cylindrical: A cylindrical shell of charge

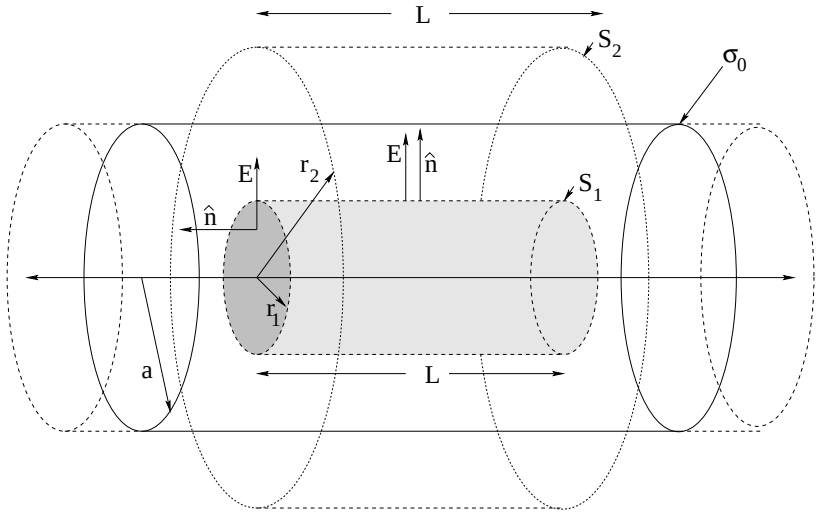


Figure 2.10: A cylindrical shell of radius a, carrying a uniform charge per unit area σ_0. Two cylindrical concentric *Gaussian surfaces* S_1 (with radius $r < a$ and S_2 (with radius $r > a$) are shown.

> 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 $\boldsymbol{r}$ and be constant in magnitude and perpendicular to the surface at all points on our Gaussian surface from the symmetry of the distribution. On the end caps the field may well vary with r, but it is *parallel* to those surfaces and therefore there is no net flux through the caps. Hence:

$$\phi_e \;=\; \oint_{S_1} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\, dA$$

$$\begin{aligned} &= \phi_{\text{caps}} + E_r \int_{\text{Cyl}} dA \\ &= 0 + E_r(2\pi r)L \end{aligned} \tag{2.64}$$

where it is presumed that everybody knows how to integrate to evalute the area of a cylindrical surface of radius r and length L *and* knows the result[3]. 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:

$$\phi_e = \oint_{S_1} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\, dA = E_r(2\pi r L) = \frac{Q_{S_1}}{\epsilon_0} = 0 \tag{2.65}$$

and solve for E_r:

$$\begin{aligned} E_r(2\pi r L) &= 0 \\ &= \frac{0}{2\pi r L} \\ E_r &= 0 \qquad \text{for } r < a \end{aligned} \tag{2.66}$$

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*:

$$\begin{aligned} \phi_e &= \oint_{S_2} \boldsymbol{E} \cdot \hat{\boldsymbol{r}}\, dA \\ &= \phi_{\text{caps}} + E_r \int_C dA \\ &= E_r(2\pi r)L \end{aligned} \tag{2.67}$$

and in fact it will *always* be this algebraic form for a cylindrical problem, to the point where we will get bored writing this line out umpty times doing homework. Don't let that stop you! Do it every time, as when you know

[3]Think of the label of a soup can. Use mental scissors to snip, snip, snip it off. Unroll it in your mind. It is $2\pi r$ long and L wide.

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 $ad\theta\ dz$):

$$\begin{aligned} Q_{S_2} &= \int_S \sigma_0\ dA = \int_0^{2\pi} d\theta \int_{-L/2}^{L/2} a\sigma_0\ dz \\ &= 2\pi a L \sigma_0 \end{aligned} \tag{2.68}$$

which we *could* have done using our heads instead of calculus, but again this way you get to see how to do a two dimensional integral that separates into two trivial one dimensional integrals.

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

$$\begin{aligned} \phi_e = E_r(2\pi r L) &= \frac{Q_{S_2}}{\epsilon_0} \\ E_r &= \frac{2\pi a L \sigma_0}{2\pi L \epsilon_0}\frac{1}{r} \\ &= \frac{\sigma_0}{\epsilon_0}\frac{a}{r} \\ &= \frac{2k\lambda_0}{r} \end{aligned} \tag{2.69}$$

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.

2.3.4 Planar: A sheet of charge

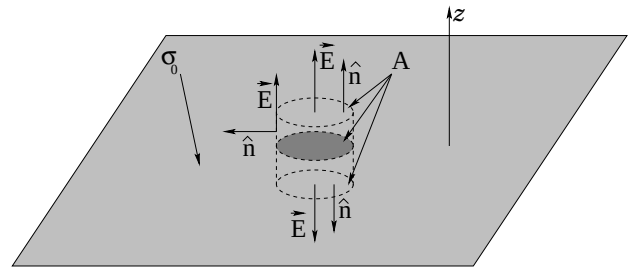


Figure 2.11: An (infinite) plane sheet of uniform charge per unit area σ_0. The Gaussian surface in this case is a simple "pillbox" symmetrically drawn so it intersects the sheet as drawn.

> 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 2.11 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 that the field is on this surface yet, but we do know that on the *end-caps* it must lie parallel to $\boldsymbol{z}$ 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:

$$\begin{aligned} \phi_e &= \oint_S \boldsymbol{E} \cdot \hat{\boldsymbol{z}}\, dA \\ &= \phi_{\text{side}} + 2E_z A \\ &= 2E_z A \end{aligned} \tag{2.70}$$

where you should note that we have *two* end caps, each of which contributes $E_z A$ to the flux.

The total charge inside this Gaussian surface is trivial:

$$Q_S = \int_A \sigma_0 \, dA = \sigma_0 A \tag{2.71}$$

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

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

$$\begin{aligned} \phi_e = 2E_z A &= \frac{Q_S}{\epsilon_0} = \frac{\sigma_0 A}{\epsilon_0} \\ E_z &= \frac{\sigma_0}{2\epsilon_0} \\ &= 2\pi k \sigma_0 \end{aligned} \tag{2.72}$$

where we note that the field is *uniform* – it doesn't depend on z, and of course it cannot depend on x and y either as every point is in the middle of an infinite plane! This last result is very important.

Note well: The parameter A (which you *made up* when you drew your Gaussian surface) *cancels* from the problem. Also note that this is exactly the result we got for the field on the axis of a disk of charge when we let the radius go to ∞. This gives us confidence that Gauss's Law *works!*

As before, in lecture your instructor will probably do a few more problems, perhaps a slab of charge of finite thickness or the field produced by *two* infinite sheets of charge, one with charge σ_0 and the other with charge $-\sigma_0$ (a model for a parallel plate capacitor that we will study in great detail shortly).

2.4 Gauss's Law and Conductors

2.4.1 Properties of Conductors

A conductor is a material that contains many "free" charges that are *bound to the material* so that they cannot easily jump from the conductor into a surrounding insulating material (where a vacuum is considered an insulator for the time being, as is air) but *free to move* within the material itself if any e.g. electrical field exerts a force on them.

In a typical conductor – for example a metal such as silver or copper – there is on average roughly one free electron *per atom* in the material. That

is order of 10^{24} electrons per mole of metal, which in turn is somewhere between 10^4 and 10^5 Coulombs! 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, or adding enough electrons so that every atom had two. 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.

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:

$$E_\perp = 4\pi k_e \sigma \tag{2.73}$$

  where σ is the charge per unit area on the surface of the conductor.

  To prove this, consider a Gaussian pillbox 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:

$$E_\perp A = 4\pi k_e Q_S = 4\pi k_e \sigma A \tag{2.74}$$

  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 $\boldsymbol{E}_\perp$ 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!

2.5 Homework for Week 2

Problem 1.

A uniform line of charge 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 $x_1 > 0$.

Problem 2.

An arc of linear charge density λ_0 and radius a is centered on the origin and subtends an angle θ_0 (which might as well start at the positive x axis and sweep counterclockwise as usual). Find the electric field at the origin.

Problem 3.

A point dipole $\boldsymbol{p}$ is located a distance r from an infinitely long line of charge with a uniform linear charge density $+\lambda_0$. Assume that the dipole is aligned with the field produced by the line charge. Determine the force acting on the dipole. Is it attracted to or repelled by the line?

Problem 4.

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 5.

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 6.

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 7.

A conducting neutral sphere of radius R is placed in a uniform electric field $\boldsymbol{E} = E_0\hat{\boldsymbol{z}}$. Using Gauss's Law and the properties of conductors in equilibrium, *draw* a representation of the electric field that results. Also indicate on the figure the qualitative distribution of charge on the surface of the conductor one expects as its charge polarizes in response to the external field.

Problem 8.

Consider three "thin" concentric conducting spherical shells with radii $R_1 < R_2 < R_3$ 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. (b) Make a table showing the net charge on the inner and outer surfaces of each conducting shell.

Problem 9.

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 $+\lambda$. Pick a point P that is in the plane of the ring but not at the center. (a) Write an expression the field produced by the small pieces of arc subtended by opposed small angles with vertex P, along the line that bisects this small angle. (b) Does this field point towards the nearest arc of the ring or the farthest arc of the ring? (c) Suppose a charge $-q$ is placed at the center of the ring (at equilibrium). Is this equilibrium stable[4]? d) Suppose the electric field dropped off like $1/r$ instead of $1/r^2$. Would you expect the electric field to vanish in the plane inside of the ring?

Problem 10.

A uniformly charged nonconducting sphere of radius a is centered on the origin and has a uniform charge density $\rho(r) = \rho_0$. (a) Show that at a point within the sphere a distance r from the center the electric field is given by:

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

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

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

in magnitude (where $\boldsymbol{b} = b\hat{\boldsymbol{x}}$). (c) Find the electric field at an arbitrary point on the x axis *outside* both spheres. Expand the result for large $x \gg a$ and keep the first *2* terms. Interpret them in terms of the expected monopolar and dipolar field of this arrangement.

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

[4] As a parenthetical aside, note that this is the problem with the ringworld described in Larry Niven's famous *Ringworld* series of science fiction novels, as gravitational attraction has the same form as the electrostatic attraction discussed in this problem.

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!

*** Problem 11.**

(A) 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 $\boldsymbol{r}_0 = (x_0, y_0, z_0)$ 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:

$$\begin{aligned}
\boldsymbol{E}(\boldsymbol{r}_0 + \Delta\boldsymbol{r}) \;=\; & \\
& \left(E_x(\boldsymbol{r}_0) + \Delta x \left.\frac{\partial E_x}{\partial x}\right|_{\boldsymbol{r}_0} + \Delta y \left.\frac{\partial E_x}{\partial y}\right|_{\boldsymbol{r}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_x}{\partial z}\right|_{\boldsymbol{r}_0} + \ldots \right) \hat{\boldsymbol{x}} + \\
& \left(E_y(\boldsymbol{r}_0) + \Delta x \left.\frac{\partial E_y}{\partial x}\right|_{\boldsymbol{r}_0} + \Delta y \left.\frac{\partial E_y}{\partial y}\right|_{\boldsymbol{r}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_y}{\partial z}\right|_{\boldsymbol{r}_0} + \ldots \right) \hat{\boldsymbol{y}} + \\
& \left(E_y(\boldsymbol{r}_0) + \Delta x \left.\frac{\partial E_z}{\partial x}\right|_{\boldsymbol{r}_0} + \Delta y \left.\frac{\partial E_z}{\partial y}\right|_{\boldsymbol{r}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_z}{\partial z}\right|_{\boldsymbol{r}_0} + \ldots \right) \hat{\boldsymbol{z}} +
\end{aligned} \tag{2.75}$$

where we only keep first order terms.

Noting that $\Delta A = \Delta x \Delta y = \Delta x \Delta z = \Delta z \Delta y$ and that $\Delta V = \Delta x \Delta y \Delta z$, show that the net electric flux *out* of this box is:

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

If we then take the differential limit and use Gauss's Law as we have thus far learned it, this becomes:

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

or

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{2.76}$$

Congratulations! You've just derived Gauss's Law in its *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). 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:

$$\Delta U(\boldsymbol{x}_0 \to \boldsymbol{x}_1) = -\int_{\boldsymbol{x}_0}^{\boldsymbol{x}_1} \boldsymbol{F} \cdot d\boldsymbol{x} \tag{3.1}$$

  where $\boldsymbol{F}$ is the total force due to all other charges.

- The vector electrostatic force can be found from the the potential energy function by taking its negative *gradient*:

$$\boldsymbol{F} = -\boldsymbol{\nabla} U \tag{3.2}$$

- For charge density distributions with "compact support" (ones we can draw a ball around, basically) we by convention define the zero of the potential energy function to be at ∞:

$$U(\boldsymbol{x}) = -\int_{\infty}^{\boldsymbol{x}} \boldsymbol{F} \cdot d\boldsymbol{x} \tag{3.3}$$

  For point charges q_1 and q_2, it is just:

$$U(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{k q_1 q_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|} \tag{3.4}$$

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

$$U_{\text{tot}} = \frac{1}{2} \sum_{i \neq j} \frac{k q_i q_j}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \tag{3.5}$$

(there is a similar integral expression for continuous charge distributions we will address later) where the 1/2 is to compensate for double counting in the sum.

- The electrostatic *potential* produced by a charge q is a one-body scalar field defined by:

$$V(\boldsymbol{x}) = \lim_{q_0 \to 0} \frac{U(\boldsymbol{x})}{q_0} \tag{3.6}$$

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

$$V(\boldsymbol{r}) = \frac{kq}{r} \tag{3.7}$$

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

$$V(\boldsymbol{x}) = -\int \boldsymbol{E} \cdot d\boldsymbol{x} + V_0 \tag{3.8}$$

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

$$V(\boldsymbol{x}) = -\int_{\infty}^{\boldsymbol{x}} \boldsymbol{E} \cdot d\boldsymbol{x} \tag{3.9}$$

and

$$\boldsymbol{E} = -\boldsymbol{\nabla} V \tag{3.10}$$

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

$$V_{\text{tot}}(\boldsymbol{x}) = \sum_i \frac{kq_i}{|\boldsymbol{x} - \boldsymbol{x}_i|} \tag{3.11}$$

or

$$V_{\text{tot}}(\boldsymbol{x}) = \int \frac{k dq_0}{|\boldsymbol{x} - \boldsymbol{x}_0|} = \int \frac{k\rho(\boldsymbol{x}_0) d^3 r_0}{|\boldsymbol{x} - \boldsymbol{x}_0|} \tag{3.12}$$

- Conductors at electrostatic equilibrium are *equipotential*. We can therefore speak of the *potential difference* between two conductors in electrostatic equilibrium where it doesn't matter what path we use to go from one conductor to the other. This also means that if we charge one isolated conductor to some potential and then connect it to another isolated conductor, charge will flow until the two conductors (now one) are at the *same* potential, a process called *charge sharing*.

- In a strong enough electric field, *dielectric breakdown* occurs and insulators "suddenly" become conductors (e.g. lightning in air). Strong fields are often induced in the vicinity of a sharp conducting point, causing a slower *corona effect* discharge that is the basis for lightning rods.

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

3.1 Electrostatic Potential Energy

The electrostatic force is *conservative*. That is, the work done moving a charge between any two points in an electrostatic field is independent of the path taken. For conservative forces we can define the *change in potential energy* to be the negative work done by the electrostatic force moving between two points:

$$\Delta U(\boldsymbol{x}_0 \to \boldsymbol{x}_1) = -\int_{\boldsymbol{x}_0}^{\boldsymbol{x}_1} \boldsymbol{F} \cdot d\boldsymbol{x} \tag{3.13}$$

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

$$\boldsymbol{F} = -\boldsymbol{\nabla} U \tag{3.14}$$

Consequently we see that we could equally well define the electrostatic potential energy in terms of an *indefinite* integral and an *arbitrary constant of integration*:

$$\Delta U(\boldsymbol{x}) = -\int \boldsymbol{F} \cdot d\boldsymbol{x} + U_0 \tag{3.15}$$

that effectively sets the point where the potential energy is zero.

By convention, for charge densities that have *compact support* – ones that one can draw a ball of finite radius (however large that radius might be) so that it *completely contains* all of the charge – we define the potential energy to be zero at ∞, just as we did for the gravitational potential energy:

$$\Delta U(\boldsymbol{x}) = -\int_{\infty}^{\boldsymbol{x}} \boldsymbol{F} \cdot d\boldsymbol{x} \tag{3.16}$$

(so that U_0 is zero, if you prefer). We remain free to choose a different zero, however, in any problem where doing so is computationally convenient.

Using the relations above, it is easy to show that the potential energy of two point charges is:

$$U = \frac{kq_1q_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|} \tag{3.17}$$

which again looks very much like that for gravity as might be expected.

One important advantage of working with the potential energy is that it is a *scalar*. To find the total potential energy of a collection of charges, we just *add it up pairwise*:

$$U_{\text{tot}} = \frac{1}{2}\sum_{i\neq j}\frac{kq_iq_j}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \tag{3.18}$$

Note that in this sum the $1 \to 2$ interaction is counted *twice*, once as q_1q_2 and once as q_2q_1. We only wish to count it once, so we divide the result by $1/2$. Another way to deal with this issue is to order the sum so that we simply never do a pair twice:

$$U_{\text{tot}} = \sum_{i<j}\frac{kq_iq_j}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \tag{3.19}$$

This stands for "sum over all q_j and all q_i such that $i < j$" which excludes all the self-energy $i = j$ terms. Good thing, too, since they are all infinite!

3.2 Potential

The good thing about potential energy is that it is a scalar and easier to evaluate than the *vector* force or field. However, it isn't terribly easy! It is still a two-body interaction term and requires us to do a nasty double sum (that becomes an even nastier double integral) when we have a large collection of charges.

A couple of weeks ago we introduced the idea of the *field* to eliminate two body computations for electric force and to give us the comfort of an apparent action-at-a-distance *cause* of the electric force. Let us do exactly the same thing here. We will define the electrostatic *potential* to be a scalar field of "potential energy per unit charge" that is the *cause* of a charged particle placed in it having a potential energy.

The formal definition of the potential is that it is the potential energy of a small test charge q_0 interacting with all the other charges that create

the potential, per unit test charge, in the limit that this small test charge vanishes:

$$V(\boldsymbol{x}) = \lim_{q_0 \to 0} \frac{U(\boldsymbol{x})}{q_0} \tag{3.20}$$

Note that this strange-seeming condition ensures that the test charge itself doesn't perturb the charge distribution that produces the potential.

The SI units for potential are:

$$1 \text{ Volt} = \frac{1 \text{ Joule}}{1 \text{ Coulomb}} \tag{3.21}$$

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

$$V(\boldsymbol{x}) = \lim_{q_0 \to 0} \frac{1}{q_0} \frac{kqq_0}{|\boldsymbol{x} - 0|} = \frac{kq}{r} \tag{3.22}$$

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

$$V(\boldsymbol{x}) = -\int \boldsymbol{E} \cdot d\boldsymbol{x} + V_0 \tag{3.23}$$

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

$$V(\boldsymbol{x}) = -\int_{\infty}^{\boldsymbol{x}} \boldsymbol{E} \cdot d\boldsymbol{x} \tag{3.24}$$

In many cases (especially when we start to treat conductors more thoroughly in later chapters) we will be interested in *potential differences*. If the field is known and well behaved, they can be easily computed by means of:

$$\Delta V(\boldsymbol{x}_1 \to \boldsymbol{x}_2) = -\int_{\boldsymbol{x}_1}^{\boldsymbol{x}_2} \boldsymbol{E} \cdot d\boldsymbol{x} \tag{3.25}$$

We can invert these relations to obtain:

$$\boldsymbol{E} = -\boldsymbol{\nabla} V \tag{3.26}$$

which in some cases will give us a relatively easy path to find the field. If the potential is relatively easy to find by (say) superposition (because it is a straight scalar sum or integral over the potentials of all the contributing

charges) then one can find the field by doing relatively easy derivatives instead of sums or integrals over vector components.

Note that this relation gives us a new way to write the strength of a field in SI units as volts per meter. Note also that there is a precise analogy between force and potential energy and field and potential. Finally, note that once we know the potential produced by a collection of *fixed* charges, we can compute the potential energy of a charge q placed in the potential *subject to the condition* that the presence of the charge in the potential does not cause significant rearrangement of the charges that create that potential as:

$$U = qV \tag{3.27}$$

This will not always be the case! In fact, if we were picky we'd say that it is almost never the case in nature, because atoms aren't "solid" objects and inevitably distort in the presence of the field of the perturbing charge. However, that doesn't really stop us from using this expression; we merely have to compute the potential energy in the *self-consistent* perturbed potential of the other charges. It does make it a bit more difficult, though.

3.3 Superposition

As we noted in the previous section, a major motivation for introducing potential is that it is a scalar quantity that we can evaluate by doing sums that don't involve the complexity of vector components or charge-charge interactions. The rule for finding the potential of a collection of charges is simple: We just add up the scalar potential of each (point-like) charge independent of all the rest!

This is once again the *superposition principle* for electrostatics, now applied to the scalar potential:

$$V_{\rm tot}(\boldsymbol{x}) = \sum_i \frac{kq_i}{|\boldsymbol{x} - \boldsymbol{x}_i|} \tag{3.28}$$

In words, the potential at a point in space is the simple (scalar) sum of the individual potentials of all the charges that contribute to that total potential.

As before, when we are working at scales where there are many many elementary point charges contributing to the potential, we can coarse grain

average. That is, we can look at a volume ΔV that is large enough to contain sufficient charge for a smooth average charge density to result that is also small enough that we can sum over it as if it is the integration volume element dV (or ditto for surface or linear distributions with elements dA and dx respectively).

Then the sum becomes:

$$\begin{aligned} V_{\rm tot}(\boldsymbol{x}) &= \int \frac{k\ dq_0}{|\boldsymbol{x}-\boldsymbol{x}_0|} \\ &= \int \frac{k\ \rho(\boldsymbol{x}_0)\ d^3r_0}{|\boldsymbol{x}-\boldsymbol{x}_0|} \qquad \text{volume} \qquad (3.29) \\ &= \int \frac{k\ \sigma(\boldsymbol{x}_0)\ d^2r_0}{|\boldsymbol{x}-\boldsymbol{x}_0|} \qquad \text{area} \qquad (3.30) \\ &= \int \frac{k\ \lambda(\boldsymbol{x}_0)\ dr_0}{|\boldsymbol{x}-\boldsymbol{x}_0|} \qquad \text{line} \qquad (3.31) \end{aligned}$$

3.3.1 Deriving or Computing the Potential

The rules above give us two distinct ways to evaluate the potential in any given problem, and we must look at the problem carefully to assess which one is best.

1. 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.

2. 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

3.4.1 Potential of a Dipole on the x-axis

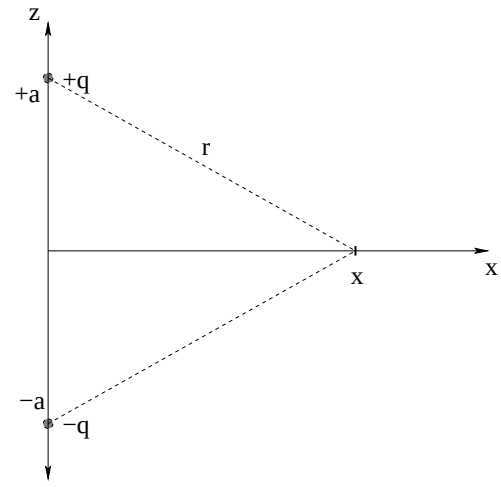


Figure 3.1: A simple dipole aligned with the z-axis.

> This is the same dipole studied in the the chapter on field. Find the *potential* at an arbitrary point on the x-axis.

This problem is deceptively simple. We know from the superposition principle that the potential is:

$$\begin{aligned} V(x) &= \sum_{i=1}^{2} \frac{k_e q_i}{r_i} \\ &= \frac{k_e q}{(x^2+a^2)^{1/2}} - \frac{k_e q}{(x^2+a^2)^{1/2}} = 0 \end{aligned} \tag{3.32}$$

This is absolutely correct – the potential of a dipole vanishes on the *entire plane* that symmetrically bisects the line connecting the charges.

The "deception" occurs when we try to compute the *field* by using $\boldsymbol{E} = -\boldsymbol{\nabla} V$. We are ever so tempted to go e.g.:

$$E_z = -\frac{dV}{dz} = -\frac{d0}{dz} = 0 \tag{3.33}$$

which is simple, easy, and *wrong!* The problem is that even though the function $V(x, y, z)$ is zero at a point that does *not* mean that its *slope* is

zero at the point! We have to use L'Hopital's Rule to evaluate a derivative at a point where its lower order derivatives or value are zero.

What this means is that we have to evaluate the function for $V(x, y, z)$ *near* but not *on* the point where the function is zero, take the desired derivative, and then let the parameter that describes that nearness go to zero. In this case, we need to find $V(x, z)$ for some *small* z (near zero), take the derivative, and let the value of z in the derivative go to zero. See if you can draw pictures to verify the following algebra, for a point $z \ll a \ll x$ above the point on the x-axis.

$$V(x,z) = \frac{k_e q}{(x^2 + (a-z)^2)^{1/2}} - \frac{k_e q}{(x^2 + (a+z)^2)^{1/2}} \tag{3.34}$$

Now we can differentiate:

$$\begin{aligned} E_z &= -\frac{d}{dz}\frac{k_e q}{(x^2 + (a-z)^2)^{1/2}} + \frac{d}{dz}\frac{k_e q}{(x^2 + (a+z)^2)^{1/2}} \\ &= -\frac{k_e q(a-z)}{(x^2 + (a-z)^2)^{3/2}} - \frac{k_e q(a+z)}{(x^2 + (a+z)^2)^{1/2}} \end{aligned} \tag{3.35}$$

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

$$\begin{aligned} E_z &= -\frac{2k_e qa}{(x^2 + a^2)^{3/2}} \\ &= -\frac{k_e p_z}{(x^2 + a^2)^{3/2}} \end{aligned} \tag{3.36}$$

Compare this to equation (1.22)! Hmmm, looks the same[1]! 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

[1] Allowing, of course, for the change in the name of the vertical axis...

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).

3.4.2 Potential of a Dipole at an Arbitrary Point in Space

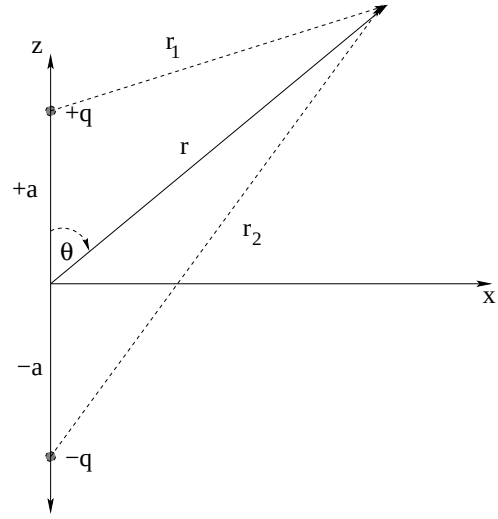


Figure 3.2: A simple dipole aligned with the z-axis, in a spherical coordinate system.

> Find the potential of this dipole at an arbitrary point $P = (r, \phi, \theta)$. Because the problem is manifestly *azimuthally symmetric* the answer cannot depend in any way on ϕ (the azimuthal/longitude coordinate), so we might as well label the point $P = (r, \theta)$ in the plane of the figure, where the answer can be azimuthally rotated by ϕ about the z-axis to any other plane without changing the form of the answer.

The potential in this problem is extremely easy to find *if you can remember the law of cosines:*

$$r_1 = r^2 + a^2 - 2ar\cos(\theta) \tag{3.37}$$

$$r_2 = r^2 + a^2 + 2ar\cos(\theta) \tag{3.38}$$

so that the potential can be read off by inspection:

$$V(r,\theta) = \frac{k_e q}{(r^2 + a^2 - 2ar\cos(\theta))^{1/2}} - \frac{k_e q}{(r^2 + a^2 + 2ar\cos(\theta))^{1/2}} \tag{3.39}$$

Of course, if you *don't* remember the law of cosines, you should visit the math chapter and learn to derive it in two or three lines so you don't ever forget it again, as we will use it fairly often and you don't want this to be an obstacle to your learning!

To find the field *now*, one can take the gradient of this exact result. However, actually taking gradients is beyond the immediate scope of this course, so just bear in mind that you *can* (and if you are a physics major, almost certainly sooner or later *will*) and otherwise forget it. Doing so isn't particularly simple in any event because of the fairly complicated denominators (although it is still much easier than finding the field directly).

Consider what happens, though, when one looks at the potential at a point $r \gg a$, so far away that the dipole looks like a "point object". To find the potential then, we must use the binomial expansion to factor out the leading r dependence and to move the complicated stuff from the denominator to the numerator (losing the square roots in the process). That is:

$$\begin{aligned}
\lim_{r\gg a} V(r,\theta) &= \frac{k_e q}{(r^2+a^2-2ar\cos(\theta))^{1/2}} - \frac{k_e q}{(r^2+a^2+2ar\cos(\theta))^{1/2}} \\
&= \frac{k_e q}{r}\left\{(1-2\frac{a}{r}\cos(\theta)+\frac{a^2}{r^2})^{-1/2} - (1+2\frac{a}{r}\cos(\theta)+\frac{a^2}{r^2})^{-1/2}\right\} \\
&= \frac{k_e q}{r}\left\{(1+\frac{a}{r}\cos(\theta)-\frac{a^2}{2r^2}+...) - (1-\frac{a}{r}\cos(\theta)-\frac{a^2}{2r^2}+...)\right\} \\
&= \frac{k_e q}{r}\left\{2\frac{a}{r}\cos(\theta)+\mathcal{O}\left(\frac{a^3}{r^3}\right)\right\} \\
&\approx \frac{k_e 2qa}{r^2}\cos(\theta) \\
&\approx \frac{k_e p_z}{r^2}\cos(\theta)
\end{aligned} \tag{3.40}$$

This *is* a very simple form and is a very important one as well! It is the *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.

3.4.3 A ring of charge

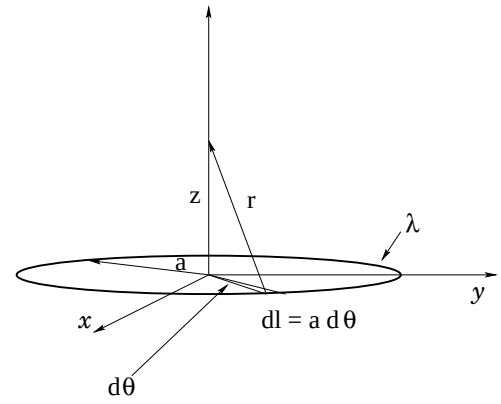


Figure 3.3: A ring of charge in the xy-plane, concentric with the z-axis.

> Suppose you are given a ring of charge with charge per unit length λ and radius a on the xy-plane concentric with the z-axis. Find the potential at an arbitrary point on the z axis.

Although there is a quick and easy answer to this problem (that will be apparent at the end, if not at the beginning) we will work through this problem in detail to illustrate the general methodology of finding a potential by integrating over a continuous distribution of charge. The steps are:

1. In suitable coordinates, define a differential "chunk" of the charge. In this problem, that would be a differential-size arc segment of the ring.

2. 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).

3. 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!

4. 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).

5. Simplify, if appropriate or required.

6. If one wishes to find the field from the potential, remember e.g.

$$E_z = -\frac{dV}{dz}$$

Beware L'Hopital's Rule! That is, if differentiating someplace that the function itself vanishes (or its functional dependence on certain coordinates vanishes) be sure that you differentiate at a general point *near* the limit point and *then* take the limit!

Let's step through this.

$$dl = a\ d\theta \tag{3.41}$$

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

$$dq = \lambda\ dl \tag{3.42}$$

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

$$dV(z) = \frac{k_e\ dq}{r} = \frac{k_e \lambda a\ d\theta}{(z^2 + a^2)^{1/2}} \tag{3.43}$$

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

$$\begin{aligned} V(z) = \int dV &= \int_0^{2\pi} \frac{k_e \lambda a\ d\theta}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e (2\pi a)\lambda}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e Q}{r} \end{aligned} \tag{3.44}$$

where we used the fact that $2\pi a\lambda = Q$, the *total charge of the ring!*

This final answer we can easily *understand* and might have even guessed without doing an integral. *All* of the charge of the ring is the *same distance* r from the point of observation, and potential depends *only* on this distance

(not on direction) so the potential is just k_e times the total charge divided by that distance.

If we do indeed try to find the electric field by differentiating this last result:

$$\begin{aligned} E_z &= -\frac{d}{dz}\frac{k_e(2\pi a)\lambda}{(z^2+a^2)^{1/2}} \\ &= \frac{k_e(2\pi a)\lambda z}{(z^2+a^2)^{3/2}} \\ &= \frac{k_e Q z}{(z^2+a^2)^{3/2}} \end{aligned} \tag{3.45}$$

Compare this to equation (2.17) above. Hmmm, looks like they are the same! However, evaluating the potential integral and then taking its derivative seems (to me, at any rate) to be *much easier* than doing the integral to find the field directly, with all of its components, and that's *before* we evaluated the E_x and E_y fields explicitly.

Note that we can exploit the insight we gained from this problem in a variety of ways to answer certain questions concerning the potential "by inspection". For example:

- A ring of charge Q a distance $R = (a^2 + z^2)^{1/2}$ from the point of observation;
- An arc of charge Q that has angular width θ and radius R, at the center of curvature;
- A hemispherical shell of charge Q with a radius R, at the center of the (hemi)sphere;
- Six charges each with charge $Q/6$ arranged in a hexagon that has a distance $2R$ between opposing corners, at the center;
- A single charge Q a distance R from the point of observation;

all produce a potential k_eQ/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.

3.4.4 Potential of a Spherical Shell of Charge

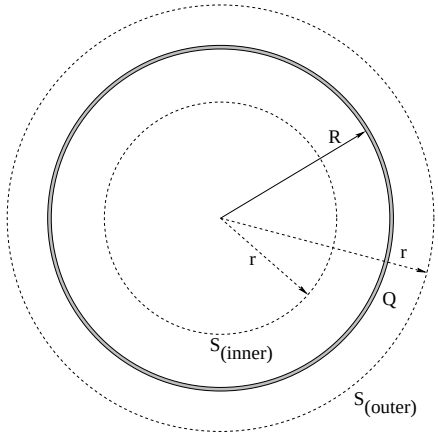


Figure 3.4: A spherical shell of charge of radius R.

> Suppose you are given a spherical shell of radius R of uniformly distributed charge Q. Find the field and the potential at all points in space.

If we want to find the potential produced by a spherical shell (or other spherical distribution of charge) and try to find it by direct integration of the potential of all the charges that make up the shell, we'll quickly discover that while it is easy to write down the integral we need to solve in some system of coordinates, it isn't so easy to *do* the integral. It's still possible – good students of calculus or students who just want a challenge can tackle it with a reasonable chance of success – but it isn't terribly easy.

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:

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

in sphere-centered spherical coordinates. We recall that the potential of any charge distribution with compact support can be found from the field by directly integrating the field according to:

$$V(\boldsymbol{r}) = -\int_{\infty}^{\boldsymbol{r}} \boldsymbol{E} \cdot d\boldsymbol{l} \tag{3.46}$$

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

$$V(\boldsymbol{r}) = -\int_{\infty}^{r} \frac{k_e Q}{r^2} dr = \frac{k_e Q}{r} \tag{3.47}$$

for all $r > R$. Inside:

$$V(\boldsymbol{r}) = -\int_{\infty}^{R} \frac{k_e Q}{r^2} dr - \int_{R}^{r} 0 \; dr = \frac{k_e Q}{R} \tag{3.48}$$

which is *constant* everywhere inside the sphere! This not only makes sense, we'll make this into a *rule*. Any volume where the electrical field vanishes has a *constant potential* – we call such a region *equipotential*. We'll talk about equipotential regions below when discussing conductors in electrostatic equilibrium (which are, as you can probably already see, equipotential).

A spherical shell of charge thus produces a potential *outside* that looks like the potential of a point charge at the origin to match its field that looks like that of a point charge at the origin. *Inside*, its potential is constant, the value it had on the shell itself coming in from the outside.

Now, a bit of warning based on my many years of teaching this class. For some of you, the first time you see a problem like this on a quiz with a region where the field is zero, the Devil is going to whisper into your ear "C'mon, dude. The field in these is zero, so the potential in there must be zero too. Put down zero and let's move on." Unfortunately, if you listen to the Devil, you'll be condemned to Physics Quiz Hell, because this would be *wrong!* Remember that the electrical field is basically the derivative of the potential. The derivative of *any constant* is zero, not just the *particular* constant whose *value* is zero.

Think of it in terms of the tops of mesas, flat mountains. Anyplace that is "flat" in potential has no field. A charge placed there doesn't gain energy moving around. But that doesn't mean that the *height* of the mesa is sea-level, or that one doesn't have to climb a steep slope from sea-level to reach the flat part. Similarly, we may have to do quite a bit of work to push a test charge from infinity to the edge of a spherical shell of charge, but once we go inside the field vanishes and we can move it anywhere without doing work. The potential inside is constant, but that constant has to reflect the *total* work done coming in from infinity (per unit charge) and is not particularly likely to be *zero*.

3.4.5 Potential of a Spherical Shell of Charge

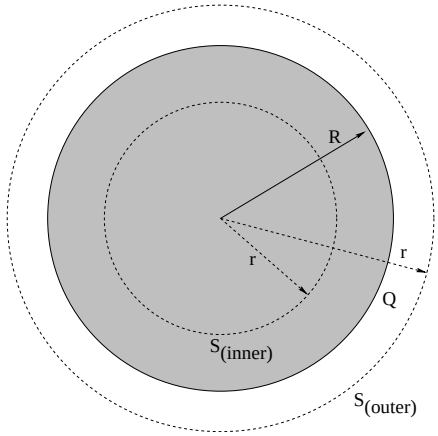


Figure 3.5: A solid sphere of uniform charge density ρ and radius R.

> Find the field *and* the potential at all points in space of a solid insulating sphere with uniform charge density ρ and radius R.

If you will recall, finding the field of a solid sphere of charge is *both* an example in the text above and was a homework assignment a couple of weeks ago – so by now you should have gone over it repeatedly and made it your own. The result was:

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

and

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

for the exterior and interior of the sphere (where we used $4\pi k_e = 1/\epsilon_0$ in the last equation just so you don't completely forget this relation as we prefer to work with k_e but one day you'll need to be able to work with ϵ_0). So just to humor me, get out paper and prove (to yourself, if nobody else) that you can still get this result, starting with Gauss's Law and *without looking*.

With the field(s) in hand, we now recapitulate the reasoning of the previous example. The distribution of charge has compact support, so we can

integrate in from infinity to find the potential (relative to infinity):

$$\begin{aligned} V(r) &= -\int_\infty^r \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_\infty^r E_{r>R} dr \\ &= -\int_\infty^r k_e Q\ r'^{-2} dr' \\ &= \frac{k_e Q}{r} \qquad r > R \end{aligned} \tag{3.49}$$

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:

$$\begin{aligned} V(r) &= -\int_\infty^r \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_\infty^R E_{r>R} dr - \int_R^r E_{r<R} dr \\ &= -\int_\infty^R k_e \left(\frac{4\pi R^3\rho}{3}\right) r'^{-2} dr' - \int_R^r k_e \left(\frac{4\pi\rho}{3}\right) r' dr' \\ &= k_e\left(\frac{4\pi R^2\rho}{3}\right) + k_e\left(\frac{2\pi\rho}{3}\right)\left\{R^2 - r^2\right\} \\ &= 2\pi k_e \rho R^2 - k_e\left(\frac{2\pi\rho}{3}\right) r^2 \qquad r < R \end{aligned} \tag{3.50}$$

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 ∞. $\boldsymbol{E}$ 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!

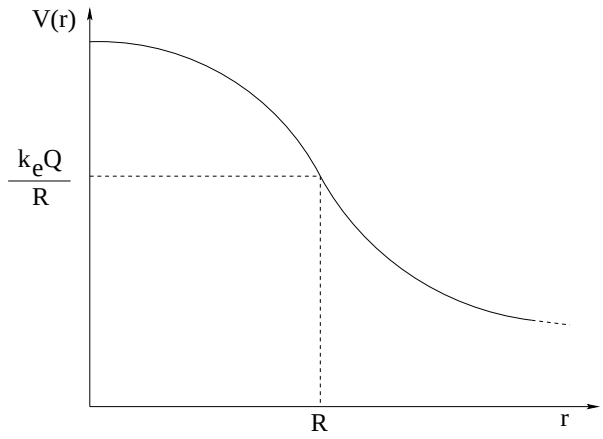


Figure 3.6: The potential produced by a uniform sphere of charge both inside and outside, as a function of r.

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.*

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

$$Q(r) = \frac{4\pi}{3}\rho r^3 \tag{3.51}$$

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

$$V(r) = \frac{k_e Q(r)}{r} = \frac{4\pi k_e}{3}\rho r^2. \tag{3.52}$$

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:

$$dW_{(\text{us})} = dU_{(\text{ball})} = V(r)dQ = V(r)\rho 4\pi r^2 dr \tag{3.53}$$

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":

$$\begin{aligned} U &= \int_0^R V(r)\rho 4\pi r^2 dr \\ &= \int_0^R \frac{4\pi k_e}{3}\rho r^2 \rho 4\pi r^2 dr \\ &= k_e \frac{16\pi^2\rho^2}{3}\int_0^R r^4 dr \\ &= k_e \frac{16\pi^2\rho^2}{3}\frac{R^5}{5} \\ &= \frac{3}{5}\frac{k_e\left(\frac{4\pi R^3}{3}\rho\right)^2}{R} = \frac{3}{5}\frac{k_e Q^2}{R} = \frac{3}{5}V(R)Q \end{aligned} \tag{3.54}$$

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 $R \to 0$, we see that $U(R \to 0) \to \infty$! If we forget the factor of 3/5, or 1/2 (which depends on the *details* of the charge distribution) and focus on the rest, we can compute a couple of extremely

interesting quantities that give us insight into nuclear physics 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 $c = 3\times10^8$ 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" [2] . 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.

[2]Wikipedia: http://www.wikipedia.org/wiki/Classical Electron Radius.

3.4.6 Potential of an Infinite Line of Charge

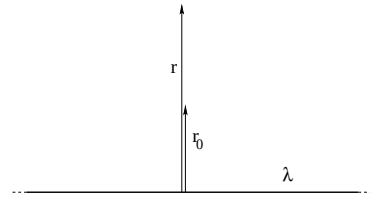


Figure 3.7: An "infinitely long" line of uniform charge density λ.

> 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:

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

in cylindrical coordinates, so that $\hat{\boldsymbol{r}}$ points directly away from the line. In fact, you should be able to show this *two ways* – using Gauss's Law (very easy) and by direct integration (much harder).

We can thus equally easily write down an expression for the potential at a distance r from the line:

$$V(r) = -\int_{\infty}^{r} \frac{2k_e\lambda}{r'}dr' = -2k_e\lambda\,(ln(r) - \ln(\infty)) = \infty - 2k_e\lambda\ \ln(r) \quad (3.55)$$

Oops. Looks like our potential is *infinite*. That's a problem...

To solve it, we compute the potential not relative to infinity but to some particular radius r_0:

$$V(r) = -\int_{r_0}^{r} \frac{2k_e\lambda}{r'}dr' = -2k_e\lambda\,(ln(r) - \ln(r_0)) = -2k_e\lambda\ \ln\left(\frac{r}{r_0}\right) \quad (3.56)$$

where we use the convenient property of natural logs: $\ln(a) + \ln(b) = \ln(ab)$ to simplify the final expression. If we let $r_0 = 1$ (in whatever units we are considering this can be further simplified to:

$$V(r) = -2k_e\lambda\ \ln(r) \quad (3.57)$$

but this *obscures the units* – recall that the argument of any function with a power series expansion e.g. ln *must be dimensionless*, so the "r" in this is the *ratio* of r in the units of choice to "1" in the unit of choice. Note well that this does not matter whenever we compute *potential difference*, which is the quantity that will be the most important one in the next chapter/week:

$$\Delta V(r_1 \to r_2) = -\int_{r_1}^{r_2} \frac{2k_e\lambda}{r'}dr' = 2k_e\lambda \ln\left(\frac{r_1}{r_2}\right) \tag{3.58}$$

where the natural log is *negative* (recall) when $r_1 < r_2$ so $r_1/r_2 < 1$. This makes *sense!* Note well that the potential *decreases* when we move *away* from the line in the direction of the field (as the potential energy decreases when we move in the direction of its associated conservative force).

On your own, show that we also get this expression if we form $\Delta V(r_1 \to r_2) = V(r_2) - V(r_1)$ using *any* of the forms for $V(r)$ given above (even the one with ∞ in it, as long as we are permitted to subtract $\infty - \infty = 0$, which of course is not necessarily or generally true but which *can* be true as the setting of the zero of the potential).

3.4.7 Potential of an Infinite Plane of Charge

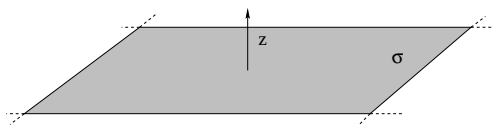


Figure 3.8: An "infinite" plane of uniform charge density σ.

> Find the field *and* the potential *relative to the plane itself* at all points in space around an infinite plane of charge. Explore the necessity of a finite reference point (where e.g. $z = 0$ is the most convenient) because the potential integrated *in* from ∞ is clearly infinite.

Using Gauss's Law (or taking the limit of e.g. a disk on its axis) you can easily show that the electric field a distance z above an infinite plane of charge with charge density σ is:

$$E_z = 2\pi k_e \sigma$$

(pointing away from the plane symmetrically on both sides) independent of z. That is, the plane of charge creates a *uniform* electric field that reaches from the plane to (in principle) ∞ without change.

If we try to evaluate the potential at a finite point z relative to ∞ we get into trouble once again because the charge distribution is non-compact:

$$V(z) = -\int_{\infty}^{z} 2\pi k_e \sigma \; dz = \infty - 2\pi k_e \sigma z \tag{3.59}$$

We feel uncomfortable with infinite quantities, so we either subtract away the infinity with a new (infinite) constant of integration, or just measure the potential difference relative to some other zero. A common, and convenient one (that leads to the same result as throwing away the infinity is $z = 0$, on the plane itself. Interestingly, this is still well defined!

$$V(z) = -\int_{0}^{z} 2\pi k_e \sigma \; dz = 0 - 2\pi k_e \sigma z = -2\pi k_e \sigma z \tag{3.60}$$

Again we will most often be interested in computing potential differences rather than potentials in the subsequent chapters, especially for non-compact charge distributions. We note that the functional variation with z is such that the potential *decreases* when one moves away from the plane; this is the most important thing to keep in mind when trying to assign or check the sign of the potential (or potential difference). The field *always* points in the direction of decreasing potential.

3.5 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 $\boldsymbol{E} = 0$ inside a conductor in electrostatic equilibrium and that $\boldsymbol{E}_{||} = 0$ at the surface, so that any electrical field immediately outside its surface must be perpendicular to the surface.

This suffices to show that conductors are *equipotential* – the potential difference between any two points in the conductor or on its surface is:

$$\Delta V = -\int_{\boldsymbol{x}_0}^{\boldsymbol{x}_1} \boldsymbol{E} \cdot d\boldsymbol{x} = 0 \tag{3.61}$$

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

$$\boldsymbol{E} = -\boldsymbol{\nabla} V_0 = 0 \tag{3.62}$$

when V_0 is any constant.

This also permits us to make an important observation. For any arrangement of (say two) isolated conductors with sufficient symmetry that we can put an arbitrary charge on either of them and not have their interaction break the symmetry of the charge's redistribution, we can compute the *potential difference* between the conducting pair as a function of the charge difference between them. This potential difference will turn out to be proportional to the charge transferred and will only otherwise depend on the *geometry* of their arrangment. In the next chapter this will be the basis of the notion of *capacitance*.

3.5.1 Charge Sharing

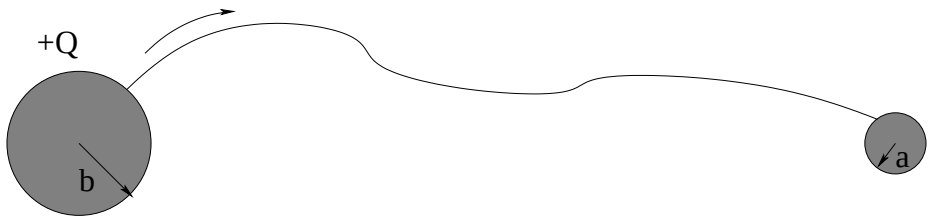


Figure 3.9: Charge sharing between two distant conductors connected by a wire. They become equipotential, with charge transferred (shared) between them to make it so.

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 3.9 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:

$$V_b = -\int_{\infty}^{b} E_r dr = \frac{kQ}{b} \tag{3.63}$$

everwhere inside sphere b while

$$V_a = 0 \tag{3.64}$$

on the other sphere. There is clearly a potential difference between the two spheres. Now imagine that we connect the two with a thin conducting wire. They form a single conductor and therefore quickly *equalize* their potentials as charge flows from b to a.

Charge is conserved. They will reach equilibrium when:

$$\frac{k(Q-q)}{b} = \frac{kq'}{b} = \frac{kq}{a} \tag{3.65}$$

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

$$\frac{q}{q'} = \frac{a}{b} \tag{3.66}$$

The smaller the sphere the smaller the fraction of charge on it, which makes sense since the *ratio* of charge to radius must be the same.

Now, however, we compute the *radial field at the surface* of the two conductors. It is:

$$E_a = \frac{kq}{a^2} \tag{3.67}$$

$$E_b = \frac{kq'}{b^2} \tag{3.68}$$

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

$$\frac{E_a}{E_b} = \frac{q}{q'}\frac{b^2}{a^2} = \frac{b}{a} \tag{3.69}$$

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

What this tells us is that the field in the vicinity of a conductor in electrostatic equilibrium at some non-zero potential is *much stronger at sharp points* than it is on smooth surfaces with a large radius of curvature. This has important consequences, as we shall see!

3.6 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, tranferring momentum to them and knocking still more electrons loose. A cascading chain reaction occurs that heats the atoms in the path of the ever increasing flow of charge and knocks still more charge loose to join that flow. In a fraction of a second, the superheated air becomes a white-hot *plasma* that conducts electricity quite well and the enormous charge difference between ground and cloud or cloud and cloud neutralizes in a burst of millions of ampere's of current. Bang! Zap! Ouch!

It is important to remember whenever working with high voltages that *few materials* are terribly good insulators against the strong fields associated with large potential differences over a short distance. That is, if you get close enough to a high voltage line it will simply arc over and electrocute you. It may well arc through a piece of glass or plastic and kill you. Wood is an insulator for ordinary voltages but conducts more than enough to kill you if you try to touch a high voltage power line with a stick.

Note also that if one approaches a conductor with a charge, one *induces* a charge on the part of the conductor nearest the charge. If that part happens to be a sharp point, the properties of charge sharing on an equipotential conductor create an *extremely strong field* in the immediate vicinity of the point. 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*.

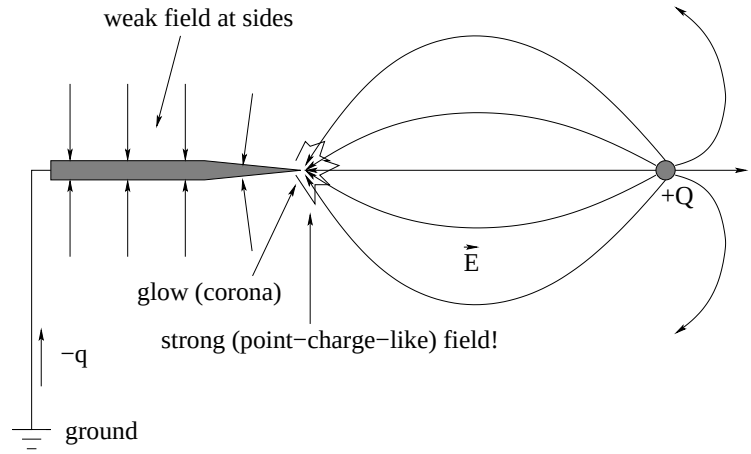


Figure 3.10: External charge $+Q$ induces a charge -q on the sharp tip of a nearby conductor. Electric fields lines leave the tip at right angles, producing a field that looks like that of a very large *point charge* which is extremely strong very close to the tip. This in turn ionizes nearby air molecules, 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.

3.7 Homework for Week 3

Problem 1.

Suppose you have charge $+q$ at position $z = a$ on the z-axis and charge $+q$ at $z = -a$. a) Write an exact expression for the eletrostatic potential of this charge arrangement at an arbitrary point (in spherical polar coordinates) $\boldsymbol{r} = (r, \theta, \phi)$. Note that the potential must be ϕ-independent because of azimuthal symmetry. You will need to recall the "law of cosines" (see the chapter on Math) to do this. b) Expand your answer to a) for $r \gg a$ and keep the lowest order surviving term. What *kind* of potential is this?

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 eletrostatic potential of the dipole at $\boldsymbol{r} = (r, \theta, \phi)$. 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 eletrostatic potential of the dipole at $\boldsymbol{r} = (r, \theta, \phi)$. 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 spheres 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 cylinders 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 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?)

*** 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 and Resistance

- Conductors *store charge* and as they do so, their *potential* (difference) *increases* relative to ground.
- If we arrange two conductors in a symmetric way and do *work* to transfer charge from one to the other (leaving behind an equal charge of the opposite sign) we call the arrangement a *capacitor* – a device for storing energy in the electrostatic field.
- The capacitance of the arrangement is defined to be:

$$C = \frac{|\Delta Q|}{|\Delta V|} \tag{4.1}$$

  or, the capacitance iS the amount of charge we can store that creates a potential difference of one volt between the conductors. Note the absolute value bars – capacitance is given as a positive quantity. The SI units of capacitance are called *farads* where:

$$1\text{F} = \frac{1 \text{ Coulomb}}{1 \text{ Volt}} \tag{4.2}$$

  A farad is an *enormous* capacitance. Typical values for capacitors in devices range from picofarads to microfarads.

  You should be able to *derive* the following quantities (from Gauss's Law, integration of potential difference, dividing into the presumed total charge):

- Parallel plate capacitor:

$$C = \frac{\epsilon_0 A}{d} \tag{4.3}$$

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

- Cylindrical capacitor:
$$C = \frac{2\pi L\epsilon_0}{lnb/a} \tag{4.4}$$
where a is the outer radius of the inner conductor, b the inner radius of the outer conductor, and L is its length (where we assume $L \gg (b-a)$).

- Spherical capacitor:
$$C = 4\pi\epsilon_0 \frac{(b-a)}{ab} \tag{4.5}$$
where a is the outer radius of the inner conductor and b the inner radius of the outer conductor.

- Energy stored in a capacitor:
$$U = \frac{1}{2}QV = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} \tag{4.6}$$
where the first form is the simplest to understand.

  One question that is very important is *where* is all this energy stored in the capacitor? The "best" answer will be: in the electric field! If we write the energy in terms of the electric field, we find that the *energy density of the electric field* is given by:
$$\eta_e = \frac{1}{2}\epsilon_0 E^2 \tag{4.7}$$

- Adding capacitors in parallel:
$$C_{\text{tot}} = C_1 + C_2 + ... \tag{4.8}$$

- Adding capacitors in series:
$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + ... \tag{4.9}$$

- Dielectrics are *insulators* that *polarize* when placed in an electric field. This builds up a surface charge that *reduces* the electric field inside the material – it *displaces* it from its usual value. For "weak fields" this reduced field is:
$$\boldsymbol{E} = \frac{\boldsymbol{E}_0}{\epsilon_r} \tag{4.10}$$

where $\boldsymbol{E}_0$ is the external field, $\boldsymbol{E}$ is the field inside the dielectric, and $\epsilon_r \geq 1$ is the *relative permeability* (also called the *dielectric constant* κ) and is characteristic of the material.

One can also describe dielectrics by shifting the *permittivity* relative to the vacuum permittivity ϵ_0 we've used so far and using it to compute the electric field inside the material:

$$\epsilon = \epsilon_r \epsilon_0 \tag{4.11}$$

- Dielectrics perform three important functions in the engineering of capacitors:

  1. They physically separate the plates (which, recall, experience a possibly strong force of attraction).
  2. 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:

     $$C = \kappa C_0 \tag{4.12}$$

     where C_0 is the capacitance without the dielectric.
  3. 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.

- 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*.

- Current is defined as:

  $$I = \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt} \tag{4.13}$$

  This is the charge per unit time flowing (for example) from one terminal of a battery to the other or from one plate of a capacitor to the other through a conducting pathway.

- Ohm's Law is:

$$\Delta V = IR \tag{4.14}$$

which can be modelled from:

$$R = \frac{\rho L}{A} = \frac{L}{\sigma A} \tag{4.15}$$

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:

$$\boldsymbol{J} = \frac{\Delta Q}{A\Delta t}\hat{\boldsymbol{n}} = \sigma \boldsymbol{E} \tag{4.16}$$

where $\boldsymbol{J}$ is the vector *current density*. From this we can see that *electric fields are not zero in a conductor carrying a current!*

- The power dissipated by a resistance carrying a current is:

$$P = VI = \frac{V^2}{R} = I^2 R \tag{4.17}$$

where the first form is the easiest to understand.

- Adding resistors in series:

$$R_{\text{tot}} = R_1 + R_2 + ... \tag{4.18}$$

- Adding resistors in parallel:

$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + ... \tag{4.19}$$

- Kirchhoff's Rules:

1. **Loop Rule:** The sum of the voltage changes around a circuit *loop* must be zero (conservation of energy).
2. **Junction Rule:** The sum of the currents flowing into a circuit *junction* must be zero (conservation of charge).

- *RC* circuits are simple loops where a capacitor is charged or discharged through a resistance. You should be able to derive that this charge/discharge is *exponential*, e.g.

$$V_C = V_0 e^{-t/RC} \tag{4.20}$$

for the simplest case, with time constant $\tau = RC$. This usually 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.

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:

$$\begin{aligned} E_r &= \frac{k_e Q}{r^2} \qquad (r > R) && (4.21)\\ &= 0 \qquad (r < R \text{ inside the conductor}) && (4.22) \end{aligned}$$

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

$$\begin{aligned} V &= -\int_\infty^r \frac{kQ}{r^2} dr \\ &= 0\frac{k_e Q}{r} \qquad (r \geq R) && (4.23) \end{aligned}$$

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

$$V = \frac{k_e Q}{R} \qquad (r < R) \tag{4.24}$$

We have seen how just *knowing* this solution for spherical shells, or the equivalent solution for cylindrical shells, can greatly improve our ability to solve problems quickly and easily by using superposition of these once-and-for-all solutions instead of trying to explicitly integrate the fields across all the different forms it might take in a problem with several conducting shells, although of course one will get the same answer either way.

Our discussion of capacitance *begins* with the observation that in this case (and the others we can solve, and other "odd" shaped conductors that we cannot) the potential of the conductor is *directly proportional to* the total charge on the conductor, and that the parameters in the potential besides the charge are k_e and things that describe its geometry, such as its physical dimensions and shape.

We could thus define a quantity we might call the "volticitance" of the conductor $\mathcal{V}$ so that (in the case of this example):

$$V = \mathcal{V}Q \tag{4.25}$$

with

$$\mathcal{V} = \frac{k_e}{R} = \frac{1}{4\pi\epsilon_0 R} \tag{4.26}$$

However, we often use conductors in particular arrangements to *store charge*. In general, we would like to be able to store a *lot* of charge on them with only a *small* potential difference. We thus seek instead a measure of the *capacity* of the conductor to store charge at any given voltage:

$$Q = CV = \left(\frac{1}{\mathcal{V}}\right) V = (4\pi\epsilon_0 R)V \tag{4.27}$$

where we have introduced the *capacitance*, the constant of proportionality that depends only on the geometry of the conductor.

To be specific, we define the *capacitance* of an arrangement of conductors used to store charge to be:

$$C = \frac{Q}{V} \tag{4.28}$$

where V is the potential difference across the arrangement as a function of the common charge Q used to create it. In the case of our example, the capacitance of an isolated conducting sphere is:

$$C = 4\pi\epsilon_0 R \tag{4.29}$$

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.

4.1.1 Parallel Plate Capacitor

In figure 4.1 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

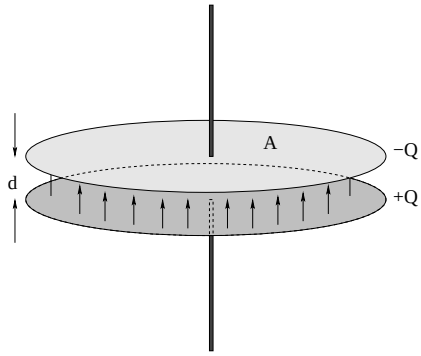


Figure 4.1: An "ideal" parallel plate capacitor of cross-sectional area A and plate separation d.

same magnitude of opposing charges on the two plates – in the static limit this is an exact result[1].

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

1. 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 substatively affect their potential difference.

2. 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.

3. 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

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

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:

$$\begin{aligned} \oint_S \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA &= 4\pi k_e Q_{\rm inS} \\ |E_z|2A &= \frac{\sigma A}{\epsilon_0} \\ E_z = \frac{\sigma}{2\epsilon_0} &= 2\pi k_e\sigma \end{aligned} \tag{4.30}$$

(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*. This we obtain a total field of:

$$E_z = 4\pi k_e\sigma = \frac{sigma}{\epsilon_0} \tag{4.31}$$

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

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"

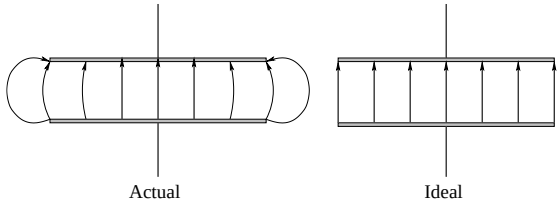


Figure 4.2: Fringe fields at the edge of an actual pair of parallel plates carrying opposite charge compared to the idealized field that vanishes sharply at the edge and is uniform in between the plates. Note that the field, and hence the potential difference, is almost identical in most of the volume between the plates.

at 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 4.2

With the fields in hand, it is but the work of a moment to compute the potential difference of the upper plate relative to the lower (or vice versa):

$$V = \Delta V = -\int_0^d E_z\, dz = -4\pi k_e \sigma d = -\frac{Qd}{\epsilon_0 A} \tag{4.32}$$

Note that the integral we computed is *negative*, which simply means that the upper plate is at a lower potential than the lower plate (consistent with the field pointing from the lower to the upper plate).

We are ready to form the capacitance. Our potential difference is negative, but when we form the capacitance we by convention make it a positive number – obviously the capacitance is symmetric and we can charge the plates in either direction, so there is no point in giving it a sign. We correspondingly form:

$$C = \frac{|Q|}{|V|} = \frac{Q}{\frac{Qd}{\epsilon_0 A}} = \frac{\epsilon_0 A}{d} \tag{4.33}$$

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!

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:

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

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

$$V = \Delta V = -\int_a^b E_r\, dr = -2k_e\lambda \ln\left(\frac{b}{a}\right) = -\frac{1}{2\pi\epsilon_0}\frac{Q}{L}\ln\left(\frac{b}{a}\right) \tag{4.34}$$

This is negative because we integrated from inside out (in the direction of the field). We could just as easily have integrated from outside in and gotten a positive potential difference. As always, the only thing that matters is that the potential must decrease when moving in the direction of the field.

The capacitance is now easy:

$$C = \frac{Q}{V} = \frac{2\pi\epsilon_0 L}{\ln\left(\frac{b}{a}\right)} \tag{4.35}$$

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.

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:

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

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

$$\begin{aligned} V &= \Delta V = -\int_b^a E_r dr \\ &= k_e Q \left\{ \frac{1}{a} - \frac{1}{b} \right\} \\ &= k_e Q \left\{ \frac{b-a}{ab} \right\} \\ &= \frac{1}{4\pi\epsilon_0} Q \left\{ \frac{b-a}{ab} \right\} \end{aligned} \tag{4.36}$$

This time I cleverly integrated from the outside in, *recognizing* that this would give me a positive potential difference as I integrate *against* the direction of the field. Now finding the capacitance is easy:

$$C = \epsilon_0 \frac{4\pi ab}{b-a} \tag{4.37}$$

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:

$$V = \frac{Q}{C} \tag{4.38}$$

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

$$\Delta W = V \Delta Q \tag{4.39}$$

The work the *blue devil* does charging up the plates is *equal* to the change in the potential energy of the charged plates[2]. We make the chunk of charge being moved differentially small, and write:

$$dU = V\, dQ = \frac{Q}{C}\, dQ \tag{4.40}$$

and can easily *integrate both sides* to find the total energy stored on the capacitor when we begin with *no* charge and charge it up to a total charge Q_0:

$$U = \int dU = \frac{1}{C}\int_0^{Q_0} Q\, dQ = \frac{1}{2}\frac{Q_0^2}{C} \tag{4.41}$$

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

$$U = \frac{1}{2}\frac{Q_0^2}{C} = \frac{1}{2}CV_0^2 = \frac{1}{2}V_0 Q_0 \tag{4.42}$$

(where note, we use $Q_0 = CV_0$ to go from the first to the second, then use it again to go to the third).

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

[2]Think of the work *you do* lifting a book over your head being equal to the *increase* in its gravitational potential energy – the work done by gravity, or the electric field in the case of the capacitor, is the opposite of the work done by you or the devil.

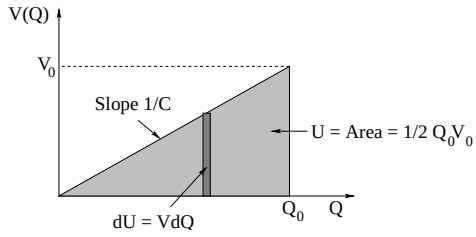


Figure 4.3: The energy as the area underneath the curve $V(Q) = Q/C$.

$dU = V\,dQ$ is just the *area* under this straight line at the particular values Q_0 and $V_0 = Q_0/C$. This, in turn, is just the area of a triangle – one half the base times the height. Which is, as you can easily see in figure 4.3, $1/2Q_0V_0$. 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 $3/5$).

4.2.1 Energy Density

A very important question to ask is: just where *is* all of this energy in the capacitor stored? We did a lot of work charging up the capacitor, and all of the work we can get back comes from charge we've stored in this way being driven *by the electric field of the charge itself* back into equilibrium as the separated charges neutralize and the field collapses. It is therefore *reasonable* to guess that the energy is stored *in the electric field we create* as we rearrange the charge in the first place.

Can we write the energy of the capacitor in terms of the field strength? Yes we can! For simplicity, we'll as usual in this chapter consider the parallel plate capacitor to see how, and then note that 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:

$$\begin{aligned} U &= \frac{1}{2}CV^2 = \frac{1}{2}\frac{\epsilon_0 A}{d}(Ed)^2 \\ &= \frac{1}{2}\epsilon_0 E^2(Ad) = \frac{1}{2}\epsilon_0 E^2(\text{Vol}) \end{aligned} \tag{4.43}$$

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:

$$\eta_e = \frac{dU}{dV} = \frac{1}{2}\epsilon_0 E^2 \tag{4.44}$$

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:

$$\begin{aligned} dU &= \eta_e dV = \frac{1}{2}\epsilon_0 E(r)^2 4\pi r^2 dr \\ U = \int dU &= \int \eta_e dV = \frac{1}{2}\epsilon_0 \int_0^\infty E(r)^2 4\pi r^2 dr \\ &= \frac{1}{2}(4\pi\epsilon_0)\left\{\int_0^R \left(\frac{k_e Q}{R^3}r\right)^2 r^2 dr + \int_R^\infty \left(\frac{k_e Q}{r^2}\right)^2 r^2 dr\right\} \\ &= \frac{1}{2}\frac{1}{k_e}k_e^2 Q^2\left\{\int_0^R \frac{r^4}{R^6}dr + \int_R^\infty \frac{1}{r^2}dr\right\} \\ &= \frac{1}{2}k_e Q^2\left\{\frac{1}{5R} + \frac{1}{R}\right\} = \frac{1}{2}k_e Q^2\frac{6}{5R} \\ &= \frac{3}{5}\frac{k_e Q^2}{R} \end{aligned} \tag{4.45}$$

exactly as we obtained at the end of Week/Chapter 3! This is a rather complicated variation in $\boldsymbol{E}$, and yet it gives us exactly the right answer. This is strong evidence that our form is general (although as noted this evidence is not proof and a proper derivation of this expression is beyond the scope of this course). You will obtain still more evidence by verifying this expression for some other arrangements of charge in your homework.

4.3 Adding Capacitors in Series and Parallel

At this point, we know how to compute the capacitance of our three "simple" geometries, and know *in principle* how to proceed for more complicated cases (although the integrals and so on may be very difficult in the general case, as always). Once we've either computed or, even better, *measured* the capacitance of a capacitor, we won't really care much what the geometry is. We can start to treat a capacitor as an "object" in its own right, and give it a *symbol* to use in designing e.g. electrical circuits. Our "standard symbol" for a capacitor will be a pair of stylized "plates" viewed edgewise, with a wire running into each plate.

Let's use this symbol (and our knowledge that $C = Q/V$) and compute the *total* capacitance of *series* and *parallel* arrangements of capacitors. We'll start with series.

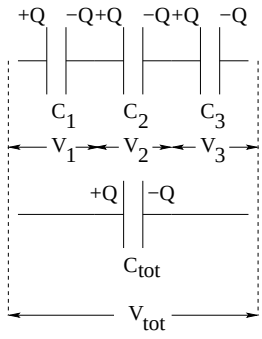


Figure 4.4: Find the total capacitance of a much of capacitors *in series*.

In figure 4.4 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 arrangment, leaving behind a charge of $-Q$). This pair of charges creates a *field* in between. However, there can *be no field* in the conducting plates and wires in the middle of the top row – they are in equilibrium! To cancel the field produced by the first plate, a charge $-Q$ is attracted to the plate facing it. But it cannot come from any part of the conducting plates or wires in between, it has to come from the surface of the next plate (leftmost of capacitor C_2) charging *it* up to $+Q$. This in turn attracts $-Q$ to the right plate of C_2, leaving a charge $+Q$ on the left plate of C_3. At this point (and you should check this) the capacitors should all be happy. Each one has a charge $\pm Q$ on it, with a field confined to live only between its plates. The field is zero inside the plates themselves and in the connecting wires. Note that all we really used in this reasoning is *charge conservation* – we couldn't create charges anywhere, only move charges around – and the idea that conductors in equilbrium can have no field inside.

Now consider the *potential differences* across each capacitor on top. Clearly the potential difference across C_1 is $V_1 = Q/C_1$, the potential difference across C_2 is $V_2 = Q/C_2$, across C_3 is $V_3 = Q/C_3$. Similarly the potential difference across our desired total capacitance is $V_{\rm tot} = Q/C_{\rm tot}$, since it has to have the *same* charge on its left plate as the arrangement on top.

Each wire between the capacitors is *equipotential*, because conductors in electrostatic equilibrium have no field inside and are thus equipotential. If we want to find the *total* potential difference across the top row of capacitors, we just have to *add up the potential difference across each capacitor*. You can think of this as doing a piecewise continuous integral across the wire at one end (get zero), the gap (pick up potential difference V_1), across the next wire (get zero), across the next capacitor's gap, (get V_2) etc. We end up with the *two* equations for the upper and lower arrangements:

$$V_{\rm tot} = V_1 + V_2 + V_3 + ... = \frac{Q}{C_2} + \frac{Q}{C_2} + \frac{Q}{C_3} + ... \tag{4.46}$$

$$V_{\rm tot} = \frac{Q}{C_{\rm tot}} \tag{4.47}$$

where the dots indicate that there was nothing special about *three* capacitors in a row – there could have been any number! We just add the potentials

across as many as we have (with the same charge on each capacitor) to get the total potential difference for the series row.

These two forms must be *equal* for equal Q on the two arrangements. That's the *definition* of the total capacitance of the upper arrangement – the equivalent single capacitor one could replace the row with and get the same potential difference for the given Q. Equating them and cancelling the common Q, we get:

$$\frac{1}{C_{\rm tot}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + ... = \sum_i \frac{1}{C_i} \tag{4.48}$$

where again the ... and final summation indicates that we just sum over as many capacitors as there are in the series row. For capacitors in series, the *reciprocal* of the total capacitance equals the sum of the *reciprocals* of the individual capacitors in series.

Why is this rule so odd? Because in series, we would get a more intuitive result by thinking of adding capacitors as if they were *volticitors*, and "volticitance" is the reciprocal of the capacitance!

Why is series addition of capacitors important and useful? Putting capacitors in series *reduces* the total capacitance (check this for yourself!) and isn't a big capacitor better than a small one? Well, yes and no. It turns out that most capacitors can only support a *finite voltage* across them before *dielectric breakdown* occurs across the intervening gap, shorting them out and burning them out. If you want to put more voltage than that maximum across a capacitor in a circuit (and don't have any rated at the desired voltage) you can put a bunch of capacitors rated at a lower voltage in series until you *can* put the desired voltage across them without exceeding the maximum for any single capacitor in the series leg. Or, you might have a bunch of big capacitors in your box and need a smaller one that wasn't in your box – adding several up in series can let you save a trip to radio shack!

So how about parallel? When several circuit elements are connected on both sides by a common conductor, the conductor on each side is *equipotential*. That means that all of the elements have the *same potential difference* across them. Note that this time I am not bothering to explicitly indicate the charge $-Q_1$ etc on the other plate of each capacitor. Recall, a capacitor is presumed to *always* have equal and opposite charges on its plates unless someone goes far out of their way to make up a problem with something

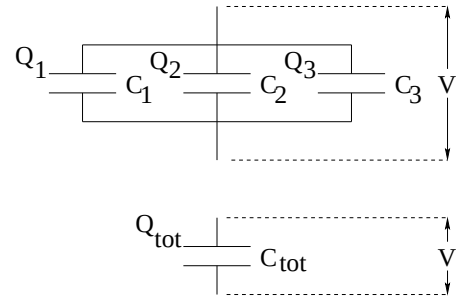


Figure 4.5: Find the total capacitance of a much of capacitors *in parallel*.

different.

In figure 4.5 *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 $C_{\rm tot}$ with the *same* voltage V across it has a charge $Q_{\rm tot} = C_{\rm tot} V$ on it. For them to be the same, the total charge store on the top arrangement has to equal that on the bottom.

This makes the problem of finding the total capacitance really easy!

$$\begin{aligned} Q_{\rm tot} &= Q_1 + Q_2 + Q_3 + ... \\ C_{\rm tot} V &= C_1 V + C_2 V + C_3 V + ... \\ C_{\rm tot} &= C_1 + C_2 + C_3 + ... = \sum_i C_i \end{aligned} \tag{4.49}$$

where we note that our rule works for *any* number of capacitors in series and write the final rule accordingly. Capacitors in parallel add!

We can understand these two rules intuitively in the following way. Capacitors in parallel increase the effective *area* where charge is stored, and hence just add. Capacitors in series increases the effective *separation* of the plates for a given area, and hence reduce the capacitance, adding reciprocally.

Before moving on, it is important to make one final observation. Capacitors (as we shall see) behave in electrical circuits the way *springs* behave in mechanical systems – they store energy and exert a restoring force on the charges that are stored that is *proportional to the charge*. Note well the analogy:

$$F_x = -k_s x \tag{4.50}$$

$$V = -\frac{1}{C}Q \tag{4.51}$$

where $1/C$ behaves like a "spring constant" and where the minus sign indicates that the potential created *opposes* the addition of more charge (we ignore this in the definition of C, but used it in the computation of U). If one computes the effective spring constant of *springs* in parallel or in series, one obtains very similar results. Springs in parallel add, with a total spring constant equal to the sum of the spring constants. Springs in series add as reciprocals, where the total spring constant is *less than the smallest* constant of the springs in the series.

Later we will learn that this analogy is nearly exact, after we discover the quantities which behave like "friction" or "drag forces" in circuits and even discover a quantity that behaves like a "mass". In the end we will find ourselves solving an equation that is identical in form to the damped, driven harmonic oscillator studied last semester, only this equation will yield the currents flowing in the circuit as a function of time. At that time it will be very fruitful to be thinking "the capacitor is like a spring" to help us understand what is going on.

4.4 Dielectrics

We have taken some care to study electric dipoles as the most common arrangement of matter that leads to an electric field, given the generally neutral character of matter. Indeed, all of the capacitors studied above can be thought of as stylized "dipoles" storing energy by separating charge. We have also observed that conductors placed in an electric field polarize and create a (mostly dipolar) arrangement of surface charge that completely cancels the electric field inside. But what of insulators? They too are made up of neutral atoms and molecules, but lack the "free charges" that carry current, as the electrons associated with each molecule prefer to stay home instead of wandering off long distances under the influence of any vagrant electric field.

To understand what a neutral atom does in the presence of an electric field, it will be very useful to have a *model* of an atom. We know that an atom consists of a tiny, massive nucleus with a charge $+Ze$ where Z is

the *atomic number* of the atom. Surrounding this nucleus is a "cloud" of Z electrons (for a total charge of $-Ze$ resulting in an electrically neutral atom), bound to the nucleus by the electrostatic force. We rather expect the neutral atom to be spherically symmetric in its distribution of charge so that there is little or no electric field outside of the charge cloud.

We still don't know *all* of Maxwell's equations, but when we do, we will be forced to confront the unpleasant truth that it is impossible for the electrons to be moving in "convenient" planetary-style classical orbits and for Maxwell's equations to be true. Of course we also don't know how to solve the associated quantum problem. Se we might as well construct the simplest possible model and hope that it provides us with some insight.

4.4.1 The Lorentz Model for an Atom

The model we will build is a to imagine the atom to consist of a pointlike nucleus surrounded by a *uniform ball* of negative charge with a total charge of $-Ze$ and a radius a (where a is around one angstrom). This is called the *Lorentz model* for the atom, and works surprisingly well – so much so that physics graduate students still use a dynamical version to understand dielectric polarization and dispersion! See figure 4.6:

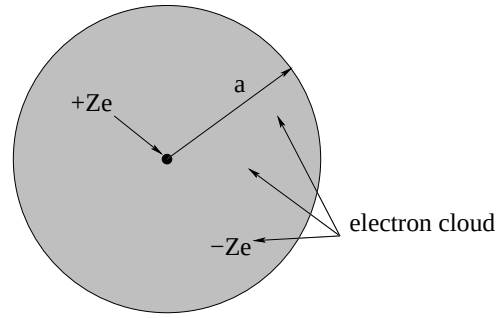


Figure 4.6: An "atom" consisting of a tiny massive nucleus surrounded by a *uniform* ball of negative charge modelling the "electron cloud".

Now we can easily *compute* what will happen when we place this atom into a "weak" electric field! We imagine that the field doesn't change the shape or size of the electron cloud but simply diplaces the nucleus away from its equilibrium position in the center to a *new* equilibrium where the force

exerted on it by the external electric field $\boldsymbol{E}_0$ balances the force on it due to the electron cloud:

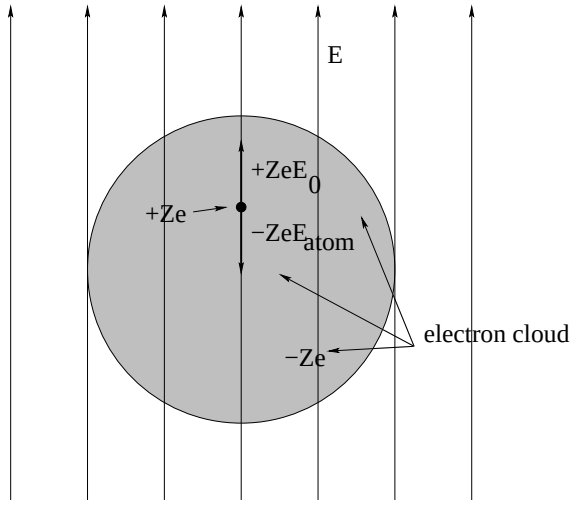


Figure 4.7: An "atom" polarized by an external electric field.

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):

$$E_{\text{atom}} = \frac{-3k_e(Ze)z}{4\pi a^3} \tag{4.52}$$

(down). Thus the forces balance when:

$$+ ZeE_0 - \frac{3k_e(Ze)^2 z_0}{4\pi a^3} = 0 \tag{4.53}$$

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

$$p_z = (Ze)z_0 = \frac{4\pi a^3}{3k_e}E_0 \tag{4.54}$$

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 $\boldsymbol{P}$. 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:

$$P_z = \frac{p_z}{8a^3} = \frac{16\pi^2\epsilon_0}{24}E_0 = \frac{2\pi^2\epsilon_0}{3}E_0 \tag{4.55}$$

where E_0 is the field in the immediate vicinity of the atom.

There was nothing special about our guestimate of a volume of $8a^3$ per atom, and of course the actual field will probably not be exactly what we compute above in the model, 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 such that:

$$\boldsymbol{P} = \chi\epsilon_0\boldsymbol{E} \tag{4.56}$$

Note well that the units of polarization are *coulombs per square meter* – those of *charge density*. It remains to find a surface for which the polarization tells us a surface charge density.

χ 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. We are only interested in the static limit of the susceptibility in an 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 of the charges that are responding. Many physicists have spent their entire careers studying quantities that amount to general susceptibilities for various materials (which can have very odd properties indeed!)

4.4.2 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!

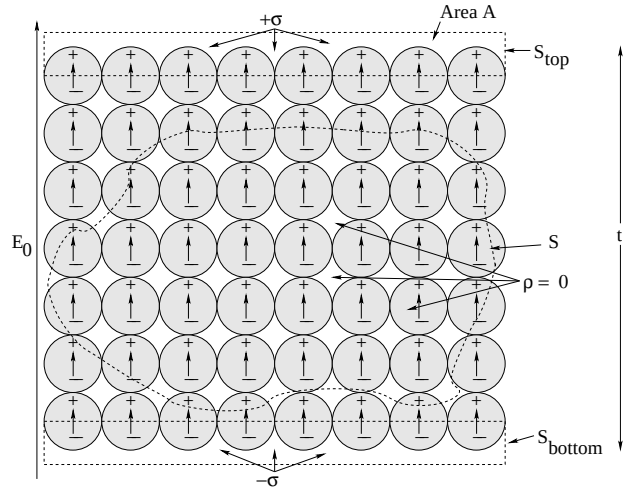


Figure 4.8: A lattice of atoms polarized by an external electric field.

In figure 4.8, we see an imaginary lattice of atoms, all polarized by an external field in the direction indicated. Note well that we've erased the *details* of even our simple model – we represent each atom as a neutral object with a small dipole moment where "some" charge is split by "some" distance by the general *process* derived and discussed in the previous section. We've drawn several possible Gaussian Surfaces inside the material.

Now let use Gauss's Law. On the *inside*, if we draw any Gaussian Surface S large enough to contain "many atoms", since the atoms are neutral the average charge inside will be *zero*[3]. Note that even where it contains an

[3]If 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

extra charge or two of either sign by splitting an atom, those charges are almost always paired with charges above or below on the neighboring atoms and the bulk remains neutral, with an average charge density $\rho \approx 0$. The interior atoms, then, do not directly modify the average field.

This is *not* true on the *surface*. If we draw a Gaussian surface $S_{\rm top}$ so that it just contains the upper half of the polarized atoms we see that it contains a nonzero positive charge; inside a similar surface $S_{\rm bottom}$ on the lower surface there is an equal and opposite negative charge. These charges make up a *surface charge layer* with a surface charge density $\pm\sigma$ that is directly proportional to E, the net field in the medium.

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:

$$p_{\rm system} = Q_{\rm surface} t = (\sigma A)t = PV = P(At) \tag{4.57}$$

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

$$\sigma = P \tag{4.58}$$

This argument is actually more general than one might suspect – if you think about it in terms of calculus you can see why it would be true for less conveniently shaped objects in a uniform field and how it might be changed to accomodate an angle between the polarization density direction at a surface and the normal to the surface there. In any event, the modifications of the field we deduce from this below are completely general and hold for arbitrary objects in nearly arbitrary fields.

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 4.9.

---

to zero as the volume gets larger.

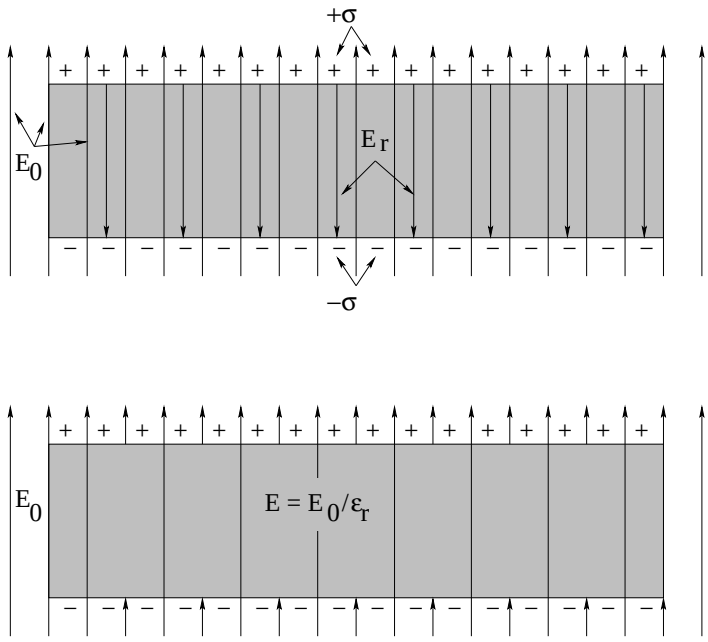


Figure 4.9: The polarized material generates a *reaction field* E_r that *opposes* the applied field and partially cancels it, making the total field in the material smaller. A dielectric material thus *reduces* the applied electric field inside the material.

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

$$E_r = \frac{\sigma}{\epsilon_0} \tag{4.59}$$

The total field is then:

$$E = E_0 - \frac{\sigma}{\epsilon_0} = E_0 - \frac{P}{\epsilon_0} = E_0 - \chi E \tag{4.60}$$

We can rearrange this into:

$$E(1 + \chi) = E_0 \tag{4.61}$$

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

$$E = \frac{E_0}{1 + \chi} = \frac{E_0}{\epsilon_r} \tag{4.62}$$

where we have introduced the *relative permittivity*

$$\epsilon_r = (1 + \chi) \tag{4.63}$$

as a dimensionless constant characteristic of the material. Note that $E \leq E_0$ because $\chi \geq 0$. This also means that $\epsilon_r \geq 1$! The electric field is *reduced* inside a dielectric – this is what the term "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 κ to represent $1 + \chi$ and call it the *dielectric constant* for the material but this usage is deprecated because in general neither ϵ_r nor κ are constant and because it encourages confusion with the *permittivity of the material*:

$$\epsilon = \epsilon_r \epsilon_0 = (1 + \chi)\epsilon_0 \tag{4.64}$$

(or equivalently, $\epsilon_r = \epsilon/\epsilon_0 = 1 + \chi$). The proper use of the permittivity in defining the electric displacement is beyond the scope of this course. We will therefore use ϵ_r in this book.

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 $\boldsymbol{E}_0$ produced by bare/isolated charges q_i 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 $\boldsymbol{E}$ by $\boldsymbol{P} = \chi\epsilon_0\boldsymbol{E}$. The net field itself is related to the applied field by $\boldsymbol{E} = \boldsymbol{E}_0/\epsilon_r$ where $\epsilon_r = 1 + \chi$.

Finally, one can equally well forget about χ and ϵ_r altogether and define the permittivity of the medium directly such that $\epsilon\boldsymbol{E} = \epsilon_0\boldsymbol{E}_0$, which can be true only if $\epsilon_r = \epsilon/\epsilon_0$ (which is, come to think of it, pretty simple). This last form suggests that the *product of the field in a material and its permittivity* should be *constant* as a field produced by any source propagates from one material to another! Perhaps we should define the *electric displacement*:

$$\boldsymbol{D} = \epsilon\boldsymbol{E} \tag{4.65}$$

This form proves to be most useful in more advanced treatments of electricity and magnetism, but is beyond the scope of this course except for being

mentioned in passing for "culture", to plant a seed or two that might flower later if you continue studying physics.

All clear now? Good...

4.4.3 Dielectrics, Bound Charge, and Capacitance

At this point you hopefully understand how a dielectric insulator is polarized by a field, how the polarization appears as a surface charge layer, how the surface charge creates a reaction field that opposes the applied field and reduces it inside the dielectric so that we can wrap *all of that up* in the simple relation:

$$E_{\text{material}} = \frac{E_0}{\epsilon_r} \tag{4.66}$$

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 4.1: Table of relative dielectric permittivities at room temperature (20° C) and some associated dielectric strenths.

So fine, so what are dielectrics *good* for? Dielectric insulators are often inserted between the plates of capacitors! Dielectrics have *three purposes* in capacitor design:

1. They mechanically separate the plates.

2. They increase the capacitance.

3. 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 0.01 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 0.01 mm (a ream of 500 sheets of printer paper is roughly 5 cm 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 0.01 mm 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. It has an area A, so $\sigma = Q/A$ and Gauss's Law tells you that the field in between the plates if there were *no* paper there would be:

$$E_0 = 4\pi\epsilon_0\sigma = \frac{\sigma}{\epsilon_0} \tag{4.67}$$

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

$$E = \frac{E_0}{\epsilon_r} = \frac{\sigma}{\epsilon_r\epsilon_0} = \frac{\sigma}{\epsilon} \tag{4.68}$$

Hmmm, seems as though the dielectric permittivity might be useful in this context, but we will restrain ourselves. Instead we will compute as usual the potential difference:

$$V = -\int_d^0 \frac{Q}{A\epsilon_r\epsilon_0} dz = \frac{Qd}{A\epsilon_r\epsilon_0} \tag{4.69}$$

and the capacitance:

$$C = \frac{Q}{V} = \epsilon_r \frac{\epsilon_0 A}{d} = \epsilon_r C_0 \tag{4.70}$$

where C_0 is the capacitance of the same geometry *without* the dielectric!

Recall that $\epsilon_r > 1$. We see that the presence of a dielectric between the plates *increases the capacitance* compare to a vacuum, or air, between the plates, *in addition* to mechanically separating the strongly attracting plates and prevenint dielectric breakdown. So what (approximately) is the capacitance of our homemade capacitor?

That's left as an exercise, a few seconds work with a calculator.

Before we move on, we need to do one final thing: relate the *free* surface charge that we put on the actual conducting plates of our parallel plate capacitor with a dielectric to the *bound* surface charge that appears on the polarized dielectric in the resulting field. We can easily do this with Gauss's Law or equivalently with our knowledge of the free field and the reaction field in terms of the surface charges.

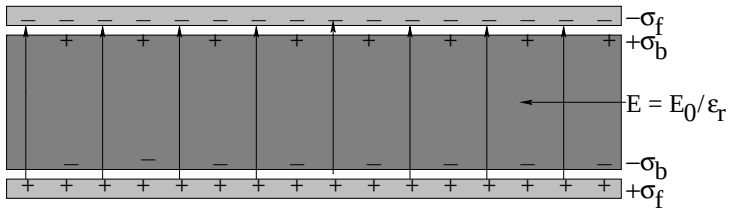


Figure 4.10: Bound and free charge in a capacitor filled with a dielectric.

In figure 4.10 we can write the field in the dielectric in two ways:

$$E = \frac{E_0}{\epsilon_r} = E_0 - E_r \tag{4.71}$$

where recall that E_r is the reaction field generated by the surface charge σ_b, which is also equal to the local polarization density at the surface. If we write out the fields E_0 and E_r in terms of the charges that produce them (basically using Gauss's law on the two surface charges), we get:

$$\frac{4\pi k_e \sigma_f}{\epsilon_r} = 4\pi k_e \sigma_f - 4\pi k_e \sigma_b \tag{4.72}$$

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

$$\frac{\sigma_f}{\epsilon_r} = \sigma_f - \sigma_b \tag{4.73}$$

or

$$\begin{aligned}\sigma_b &= \left(1 - \frac{1}{\epsilon_r}\right)\sigma_f \\ &= \left(\frac{\epsilon_r - 1}{\epsilon_r}\right)\sigma_f \\ &= \left(\frac{-\chi}{1+\chi}\right)\sigma_f \end{aligned} \tag{4.74}$$

where the last form is in terms of the material's susceptibility instead of the more commonly used ϵ_r.

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. We could show this (if our lives depended on it) by considering the derivation above as valid for differentially small volumes and using some calculus to deal with the variation, or you can take my word for it for now and (possibly) prove it later, in a more advanced class. You'll have homework problems that require you to deal with e.g. dielectrics in the space between the shells of a cylindrical or spherical capacitor, and you'll need to know this then.

As a last remark, consider field energy density inside a dielectric. If we recapitulate the argument for field energy density for a parallel plate capacitor filled with a dielectric, we get:

$$U = \frac{1}{2}CV^2 = \frac{1}{2}\frac{\epsilon_r \epsilon_0 A}{d}(Ed)^2 \tag{4.75}$$

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

$$\eta_e = \frac{dU}{dV} = \frac{1}{2}\epsilon E^2 \tag{4.76}$$

where $\epsilon = \epsilon_r \epsilon_0$ is the dielectric permittivity of the material. This is the correct form of the energy density to use inside a dielectric material.

This is all we need to know about dielectrics, although the problems will challenge you with half-filled capacitors and the like to make sure you understand it will enough to be able to use it.

4.5 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*[4].

4.5.1 Chemical Batteries

It is probably instructive to look at the actual chemical reaction associated with at least one *specific* kind of battery, even though one can make a cell out two different kinds of almost *any* metal stuck into an electrolyte solution (e.g. an acid). So let's look at the two reactions associated with a lead-acid battery, the kind you probably have in your car.

A lead-acid battery consists of two plates. The anode (positive pole) is made out of ordinary lead. The cathode (negative pole) is made of lead coated with lead oxide. Both are immersed in a solution of water and sul-

[4]Technically, a single device that generates a voltage in this way is called a *cell* – a *battery* is composed of several cells – but we'll just call anything that generates electricity a battery because nobody speaks of "flashlight cells" when they go to the store to get a pack of D's, they say "I'm going to get some batteries for the flashlight".

phuric acid. At the anode [5] :

$$Pb + HSO_4 \rightarrow PbSO_4 + H^+ + 2e^-$$

while at the cathode:

$$PbO_2 + HSO_4 + 3H^+ + 2e^- \rightarrow PbSO_4 + 2H_2O$$

The electrolyte provides both the (ionized) sulphuric acid required at both ends and a conducting pathway for the electrons to be transported from the anode to the cathode. Energy is released by this reaction; the end products are *more* stable than the original ones so the reaction is *favored*.

However, once a few atoms in the anode have given up their electrons and they've been pulled over to the cathode, the reaction stops! The poles are then *charged up* and it costs too much work to remove any more electrons, more than one *gains* in the chemical reaction. The anode is then charged up *positively* (as an electron *donor* to the reaction in the battery itself) while the cathode is charged up *negatively* (having received the electrons). The top and bottom plates behave *just like the plates of a capacitor* and maintain an electrical potential difference of around 2 volts (per *cell* in a *battery* of six cells, in a typical twelve volt battery in a car) between them that just balances the chemical potential of the arrangement.

There is, however, an important difference. If one provides a *conducting pathway* between the anode and the cathode *outside* of the solution, then the negative charge surplus on the cathode can flow *back* over to the anode and participate in another reaction, then another, then another. Charge continues to be driven in this way until all of the lead and lead oxide is converted into lead sulphate and water. For every mole of lead converted into lead sulphate, two moles of electrons have to move from cathode to anode. That is $1.2 \times 10^{24}/1.6 \times 10^{19} = 0.75 \times 10^5$ *Coulombs* of charge, enough to drive an Ampere of current (one Coulomb/second) for around a day. A mole of lead is around 207 grams, which weighs around a half a pound. Allowing for the electrolyte and sulphuric acid, roughly a pound of battery will drive a load of two watts (one ampere at two volts) for just

[5] Wikipedia: http://www.wikipedia.org/wiki/Lead-acid battery. There are more complete ways of writing out the chemical reaction that show more of what is going on with the water in all of this, but this is sufficient. Either way, you are of course encouraged to visit the link and read more about it.

under a day (where we'll work out energy relations below to justify this in a moment).

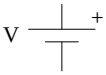
A second advantage of this particular battery is that it is *rechargable*. If one simply places a voltage across the cell that exceeds its terminal voltage, charge flows *the other way*, reversing the reaction and turning lead sulphate back into lead or lead oxide. By careful design, one can charge and discharge the battery many times before too much lead sulphate falls off of the electrodes or crystalizes out across the space in between the terminals and shorts out the batter, at which time the battery must be remanufactured (to avoid dumping toxic lead into the environment).

Vehicle batteries, of course, weight many pounds – as many as fifty or sixty – and have six cells, and therefore can drive bigger currents at higher voltages, currents that can easily be large enough to be dangerous. In fact, a car battery [6] , 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.

[6]Internet: 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.

4.5.2 The Symbol for a Battery

All of this is too complicated for intro physics, of course. We want to start by idealizing a battery and replacing it in all circuits we consider with a single simple symbol. The symbol we will use is 

, where V is the nominal potential difference maintained by the battery between its terminals (its "pole 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 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 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. 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.

Before we move on to resistance, it is worth pointing out that battery physics and engineering are *important* in our society, and becoming *more important* as we move in the direction of renewable energy sources, hybrid or flat-out electric cars, rechargable electronic devices galore and more. One of the biggest obstacles to solar or wind generated power is the difficulty of storing power generated when the sun is high and bright or when the wind blows for use at night or on a calm day. It could easily require hundreds of pounds of lead-acid batteries *per person* just to store the power needed for a

single night from sunlight collected during the say. The inventor 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, batteries have many problems: They are bulkly, massive, they get hot while operating, they are made with toxic materials, they are difficult to dispose of, they wear out, they can explode if overdriven and they tend to be expensive!

4.6 Resistance and Ohm's Law

Fine, so now we have a battery. We place a chunk of conducting matter between the poles/terminals of the battery, and what happens? Well, *current flows*, that's what happens! We have created a situation where a conductor is *not* in electrostatic equilibrium, and *charge moves in time* through the conductor in response to the force created by the battery, with *energy released* in the process. This is actuall fine, and we might even say, it's about *time* that we got out of statics (which are kind of boring, as not much happens, right?) and into *dynamics*, where things happen. All we need, then, is to come up with a model for what goes on inside the conductor as the current flows, and we can start to analyze dynamical electrical systems once again, which has to be more interesting than just thinking about a charged capacitor sitting around all do doing nothing much but just storing charge.

A microscopic picture, of course, begins with atoms, each with a heavy nucleus and surrounded by electrons, arranged in some sort of solid lattice, with some of the electrons "free" to move within the lattice. Free to move, however, is not the same thing as non-interacting. Electrons that move through the lattice interact with the lattice and transfer their momentum to the lattice so that (in equilibrium) their average velocity is zero. The lattice therefore exerts a kind of *drag force* on the electrons that brings them back to equilibrium.

The *simplest* model for conduction of electrons through a material that "resists" their motion via a drag force caused by the collision of the moving electrons with each other and the underlying atoms in the lattice is one with a *linear drag force* – one that is proportional to the average velocity of transport of the electrons through the resistive lattice. If the electrons are

being pushed through the conductor by some constant force, then, they'll arrive quickly at a *terminal velocity* that is proportional to that force, where the forces balance.

4.6.1 A Simple Linear Conduction Model

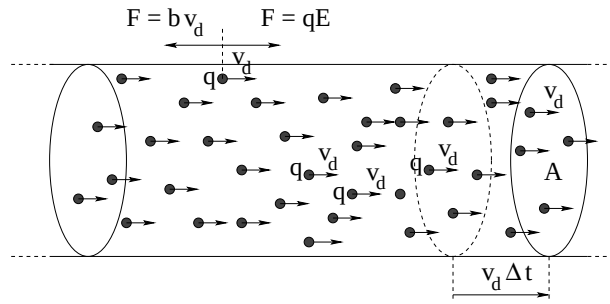


Figure 4.11: The simple linear model for conduction in a resistive lattice.

In figure (4.11) we see a model for a conducting wire. This wire has a cross-sectional area of A and contains n charges per unit volume, each with charge q. An electric field is created within the wire by a *battery* (not shown) that exerts a force to the right of $F = qE$. The wire resists the flow with a "drag force" bv_d to the left, where v_d is the so-called "drift velocity" that is the average terminal velocity of charges in the conductor. Mind you, in a typical normal metal our charge carriers are electrons and all of the vectors are reversed for a current and field that still go from left to right.

We are interested in computing the *current*: the *charge per unit time* that passes any point on the wire under the influence of the force created by the battery (or other source of potential difference across the wire). From the picture we can see that *all of the charge* ΔQ in the volume between the dashed circle and the circle at the right passes through the cross-sectional area A perpendicular to the direction of motion in a time Δt. So how much is that?

$$\Delta Q = nqv_d A\Delta t \tag{4.77}$$

which we read as "the number of charge carriers per unit volume times the charge per carrier times the volume". This means that the total charge per unit time is:

$$I = \frac{\Delta Q}{\Delta t} \approx \frac{dQ}{dt} = nqv_d A \tag{4.78}$$

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

$$1 \text{ Ampere} = \frac{1 \text{ Coulomb}}{1 \text{ Second}} \tag{4.79}$$

The result $I = nqv_dA$ 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.

4.6.2 Current Density and Charge Conservation

Note well that in the picture above, we determine the current that passes a point in the wire by evaluating how much charge passes through the *surface* perpendicular to the charge flow that passes through the point! This picture should remind you of something – it is very similar to the pictures we used to talk about *electric flux*.

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. Using arguments identical to those we worked out in our discussion of electric flux, if we want the current through a surface perpendicular to the direction of motion of the charge to be the same as the current through a second surface cut through the wire that touches the same point but is tipped at an angle θ relative to the direction of the current, the area A increases to the area $A' = A/\cos(\theta)$. 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:

$$I = nqAv_d = nq\frac{A}{\cos(\theta)}\cos(\theta) = nqA'\cos(\theta) \tag{4.80}$$

We can get the cosine out of a dot product between the local *direction* of $\vec{v}_d$ and $\hat{\boldsymbol{n}}$, a normal to the surface A or A':

$$I = nqA\boldsymbol{v}_d \cdot \hat{\boldsymbol{n}} = nqA'\boldsymbol{v}_d \cdot \hat{\boldsymbol{n}}' \tag{4.81}$$

Finally, to make this more general we can allow *curved* surfaces and flows of charge that are not all parallel, and add up the current that flows

through each tiny differential chunk of area on a completely arbitrary surface cut through the conductor:

$$I_C = \int_{S/C} nq\boldsymbol{v}_d \cdot \hat{\boldsymbol{n}} dA = \int_{S/C} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA \tag{4.82}$$

where S/C is read "through the surface S bounded by the closed curve C

$$\boldsymbol{J} = nq\boldsymbol{v}_d \tag{4.83}$$

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.*

Now suppose that one has a single curve C and two open surfaces that are bounded by it, say S_1 that cuts straight across the wire and S_2 that is ballooned out so that it resembles a fishing net, where $S_1 + S_2$ between them form a *closed* surface S containing a volume V in between them. *If* current is flowing in a "steady state" way, the current through these two surfaces must be equal – the current through the first must equal the current through the second. However, if we put e.g. a capacitor plate in between the two surfaces, current may not be flowing in a steady state way – current may be *building up* inside the *closed* surface S. In that case the difference between the current through S_1 and the current throught S_2 is the rate at which charge builds up inside V:

$$\int_{S_1} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA - \int_{S_2} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA = \frac{d}{dt} \int_{V/S} \rho_e dV \tag{4.84}$$

We can get rid of the relative minus sign by changing one of the two normal's so that it doesn't point in the left-to-right direction through the surface. If we make the normal the *outward* directed normal for the closed surface $S = S_1 + S_2$, that swaps the direction of the normal through S_1, so:

$$-\int_S \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA = \frac{d}{dt} \int_{V/S} \rho_e dV \tag{4.85}$$

which we rearrange as:

$$\int_S \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA + \frac{d}{dt} \int_{V/S} \rho_e dV = 0 \tag{4.86}$$

This equation is *very important!* It is, in fact, a *law of nature*, based on substantial empirical evidence. It is the *law of charge conservation* written

in mathematical form. Basically, it says that the amount of charge inside any volume bounded by a closed surface can only decrease (increase) if charge flows *out (or in) through the surface!* The net charge inside cannot just poof into or out of existence, it has to get there by coming in from outside[7].

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 this law and Gauss's Law and convert them to vector differential form. We leave the algebra for the conversion to then (although you may have done it in the starred homework problem in the Gauss's Law chapter earlier) but put down the result here for completeness. The law of charge conservation in differential form is:

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} + \frac{\partial \rho_e}{\partial t} = 0. \tag{4.87}$$

Again, this section is *enormously important* for things we will learn later. In fact, 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.

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. You now have all the information you need to do so but, of course, the equation that needs to be repaired. When you cover it, your instructor may point it out and suggest that you give it a try.

4.6.3 Ohm's Law

At last we are set to deduce Ohm's Law. In our simple conduction model with its linear resistive "drag" force, we noted that the (terminal) drift

[7]There is another way charges can appear inside the box that doesn't violate this law – they *can* be created or destroyed a *pair at a time* in such a way that the *net* charge of the pairs remains zero. This actually happens in high energy quantum mechanical collisions – making it beyond the scope of this course – but the creation of a positron-electron pair does not violate *net* charge conservation.

velocity v_d had to be proportional to the applied electric field E. We have just seen that the magnitude of the current density is proportional in turn to v_d. We can wrap all the constants of proportionality – which include b, n, q – into a single parameter called the *resistivity*. Unfortunately the common symbol for resistivity is ρ, which you can easily confuse with the charge density. I've tried pretty hard to label the latter ρ_e, read "the density of ELECTRIC charge" and will continue to do so whereever there is any chance of confusing the two. We expect the velocity/current density to go *down* when the resistivity of the material goes *up*, so:

$$\boldsymbol{J} = \frac{1}{\rho}\boldsymbol{E} \tag{4.88}$$

This equation is sometimes written in terms of the reciprocal of the resistivity, the *conductivity* σ (which once again collides with prior usage for the surface charge density, sorry):

$$\boldsymbol{J} = \sigma\boldsymbol{E} \tag{4.89}$$

with $\sigma = 1/\rho$.

The resistivity is a characteristic of the material of the conductor in question, and depends on many things. Its most important dependence is probably upon temperature – resistivities of most materials vary approximately linearly with temperature, increasing as the (absolute) temperature increases, but the variation is typically *slow* and can be considered nearly constant over the small range of temperatures we typically live in, but this will definitely matter if one is designing circuits that have to function across a wide range of temperatures. It also varies with pressure and other related paramters. In *this* class we won't spend a lot of time or energy thinking about this weak variation – I will simply link in Wikipedia: http://www.wikipedia.org/wiki/resistivity so you can read about it in far more detail than I can easily fit in here, complete with a nice table of temperature coefficients and a bit of theoretical explanation.

Consider a uniform conductor with resistivity ρ, length L, and cross-sectional area A. We can rearrange the equation above as:

$$\boldsymbol{E} = \rho\boldsymbol{J} \tag{4.90}$$

The electric field and current density inside of this volume are uniform (all of the charges must move to the right at the same speed or charge would

build up somewhere in the volume). If we take the flux through a *normal* cross-section of both sides we get:

$$EA = \rho JA = \rho I \tag{4.91}$$

which we can rearrange as:

$$E = \frac{\rho}{A} I \tag{4.92}$$

If we integrate both sides a second time in the direction $d\boldsymbol{l}$ from one end of the conductor to the other in the direction of the current, we get:

$$\Delta V = EL = \frac{\rho L}{A} I \tag{4.93}$$

where ΔV is the amount the electric potential *decreases* going from one side of the conductor to the other *in the direction of the field/current.*

Finally, we drop the Δ – it is still there, but will always be *assumed* to be the potential difference across any resistor, to simplify the algebra a tiny bit as we did when discussing capacitors – and define a new quantity, the *resistance* of *this particular geometry of conducting material*, R:

$$V = RI \tag{4.94}$$

where

$$R = \rho \frac{L}{A} \tag{4.95}$$

This 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,

$$1 \text{ Ohm} = frac\text{Joule} - \text{SecondCoulomb}^2 \tag{4.96}$$

Note well that the units of capacitance were coulombs squared per joule, so the units of R times C are *seconds* – this will be important to us later.

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 analagous to the current of charge. The resistance of the pipe is analagous 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.

It 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 determined by the water pressure.

In this anology, 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.

4.7 Resistances in Series and Parallel

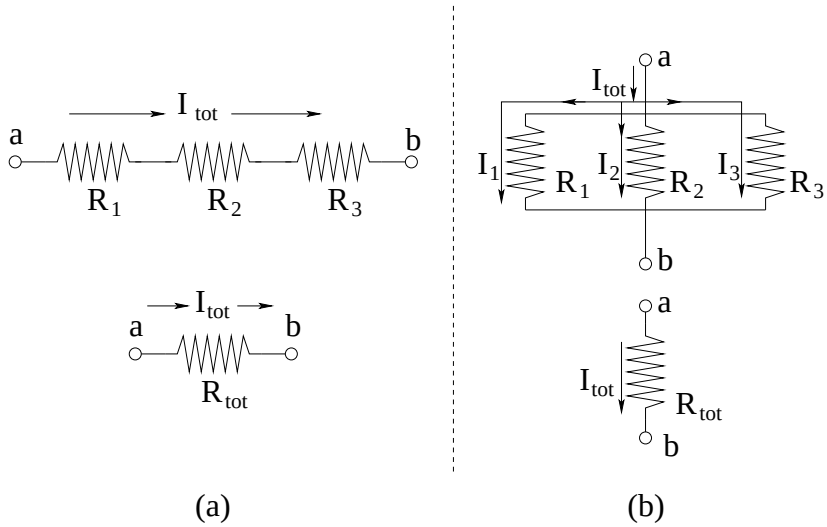


Figure 4.12: Three resistors R_1, R_2, R_3 arranged in *series* (left, (a)) and *parallel* (right, (b)), along with the equivalent/total resistances of each one portrayed below. In both cases the total resistance is "equivalent" when applying a voltage V_{ab} across the a and b contacts produces the *same total current* $I_{\rm tot}$ *in the top and bottom figure.*

In this section we indicate how to add resistances in series or in parallel in order to determine a single *equivalent* resistance that would permit the same current to flow given the same voltage across the arrangement. The algebra is simple.

4.7.1 Series

Suppose we apply a fixed voltage V_{ab} across the contacts in the upper (a) diagram. This produces some current $I_{\rm tot}$ in the *single* (serial) line of resistors. Since charge is conserved and there is nowhere for it to go but through the resistors, this same current passes through each resistor in turn. We can

thus use Ohm's Law to determine the voltage drop across *each* resistor in terms of this total current:

$$V_1 = I_{\rm tot} R_1 \tag{4.97}$$
$$V_2 = I_{\rm tot} R_2 \tag{4.98}$$
$$V_3 = I_{\rm tot} R_3 \tag{4.99}$$

Obviously the total voltage V_{ab} is given by:

$$V_{ab} = V_1 + V_2 + V_3 = I_{\rm tot}(R_1 + R_2 + R_3) \tag{4.100}$$

If we look at the lower (a) diagram, Ohm's Law yields:

$$V_{ab} = I_{\rm tot} R_{\rm tot} \tag{4.101}$$

Equating and cancelling the common $I_{\rm tot}$, we get:

$$R_{\rm tot} = R_1 + R_2 + R_3 \tag{4.102}$$

There was nothing "special" about having only three resistors. We could have had, four, five, or N resistors in series and we'd simply have more terms in a general equation:

$$V_{ab} = \sum_{i=1}^{N} I_{\rm tot} R_i = I_{\rm tot} \sum_{i=1}^{N} R_i = I_{\rm tot} R_{\rm tot} \tag{4.103}$$

so that *in general* the rule for the addition of N resistors in series is:

$$R_{\rm tot} = R_1 + R_2 + ... + R_N = \sum_{i=1}^{N} R_i \tag{4.104}$$

4.7.2 Parallel

In the case of resistances in parallel, we have the *same* voltage V_{ab} applied across all of the resistors in parallel. If we look at the upper (b) figure, we can use Ohm's Law to evaluate the *current* through each resistor, given a common voltage V_{ab} across them:

$$I_1 = \frac{V_{ab}}{R_1} \tag{4.105}$$
$$I_2 = \frac{V_{ab}}{R_2} \tag{4.106}$$
$$I_3 = \frac{V_{ab}}{R_3} \tag{4.107}$$

Now, consider the total current $I_{\rm tot}$ flowing into the arrangement from point a. Charge is conserved, so that all of the charge that flows into the first junction connecting the three independent conducting pathways through the resistors *must flow out of it* and into the three resistors. From this we conclude that:

$$I_{\rm tot} = I_1 + I_2 + I_3 = \frac{V_{ab}}{R_1} + \frac{V_{ab}}{R_2} + \frac{V_{ab}}{R_3} = V_{ab}\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) \tag{4.108}$$

As before in the lower (b) figure we have:

$$I_{\rm tot} = \frac{V_{ab}}{R_{\rm tot}} \tag{4.109}$$

and when we equate these two forms and cancel the common V_{ab} we get:

$$\frac{1}{R_{\rm tot}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \tag{4.110}$$

There is nothing special about three resistors, and once again we can easily generalize this argument to N resistors as:

$$\frac{1}{R_{\rm tot}} = \frac{1}{R_1} + \frac{1}{R_2} + ... + \frac{1}{R_N} = \sum_{i=1}^{N} \frac{1}{R_i} \tag{4.111}$$

We conclude that the total resistance of several resistors in series is the simple sum of the individual resistances, while the *reciprocal* of the total resistance of serveral resistors in parallel is the sum of the *reciprocals* of the individual resistances. This is the exact opposite of the rules for summing capacitances in seris and parallel.

4.8 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

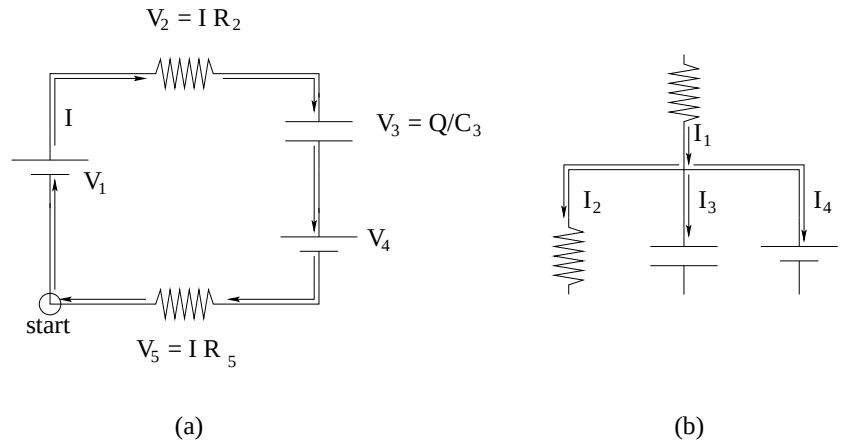


Figure 4.13: (a) A single "generic" circuit loop; (b) A single "generic" circuit junction.

wire. This is an idea related to *energy conservation*. In studying parallel capacitors and and parallel resistors, we used the idea that the total charge moving around in these circuits must be conserved to track its distribution over time whether or not it is actually moving.

These two rules (which we will derive and discuss below) are known as Kirchhoff's Rules [8] .

4.8.1 Kirchhoff's Loop Rule

Consider the generic *circuit loop* in figure 4.13 (a) above. The particular devices in this loop are not too important – I drew a fairly arbitrary mix of the three devices we are aware of so far, but later we will learn about still more devices we might want to put into a circuit to do some startlingly useful things.

Let us imagine that we watch a charge $+q$ moving around this circuit loop in the direction of the current beginning at the (arbtrary) point "start". As it goes across each potential $V_1, V_2, ...$ the energy of the charge goes up, goes down, goes up, goes down. By the time it gets back to the start position,

[8]Wikipedia: http://www.wikipedia.org/wiki/Kirchhoff's Circuit Laws.

its potential energy has changed by:

$$\Delta U = qV_1 + qV_2 + qV_3 + qV_4 + qV_5 = q\sum_i V_i \tag{4.112}$$

If $\Delta U \neq 0$, then the charge gets back to its starting point with a *different energy* than the one it started with! Its kinetic energy will have changed!

However this is *almost* impossible. Electrons in particular, as fermions, are nearly *completely incompressible* in a wire. This means that the current in any line segment is the same at all points in the segment. Changes in the electric field that *produces* the current at all points in the conductor propagate nearly *instantaneously* throughout the entire loop, because the speed of light is very large compared to the size of the loop. As potentials across the elements in the circuit vary, the current adjusts almost instantaneously. Consequently within a *very* tiny margin associated with this propagation time, the net energy gain or loss of a charge in a pass around the circuit loop must be *zero!*

This means that:

$$\sum_i^{\text{loop}} V_i = 0 \tag{4.113}$$

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.

4.8.2 Kirchhoff's Junction Rule

4.9 RC Circuits

4.10 Homework for Week 4

Problem 1.

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 $C = \frac{\epsilon_0 A}{d}$ where A is the area of the cylinder/sphere and $d = b - a \ll a$ ("small" separation).

Problem 2.

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

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

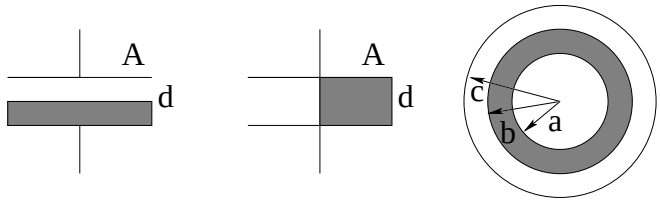
for all three geometries (where the integral is over the volume $\mathcal{V}$ between the plates); and b) $C \approx \frac{\epsilon_0 A}{d}$ for the spherical and the cylindrical capacitor, where A is the area of the plates and d is their separation. You will need to use $\ln(1 + x) \approx x + \mathcal{O}(x^2)$... to do the cylinder.

Problem 3.

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 dielectric constant κ. 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 4.

Find the capacitance of the following arrangements:



where the first two are parallel plate capacitors *half-filled* with a diectric material with dielectric constant κ as shown, and the third is a spherical capacitor patially-filled with the same dielectric as shown.

Problem 5.

Derive the rules for adding parallel and series capacitance:

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

and

$$C_{\text{tot}} = C_1 + C_2 + C_3 + ... \qquad \text{(parallel)}$$

Then derive the rules for adding parallel and series resistance:

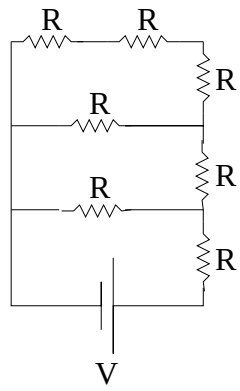
$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + ... \qquad \text{(parallel)}$$

and

$$R_{\text{tot}} = R_1 + R_2 + R_3 + ... \qquad \text{(series)}$$

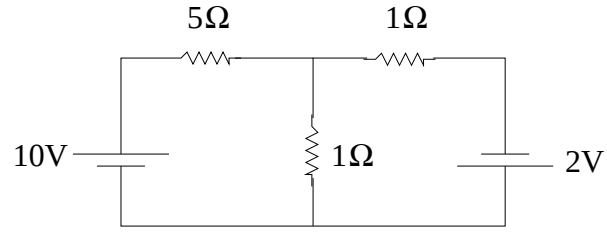
Compare and contrast the two results.

Problem 6.



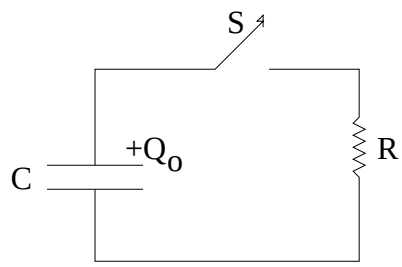
Find the current through each resistor with a voltage V is placed across the resistance network as shown to the left. Note that all of the resistances R are equal. You'll basically need to use the series and parallel rules for adding resistances several times, as well as Ohm's Law and Kirchhoff's junction rule. (Hint: You may find it useful to imagine $V = 18$ volts and $R = 1$ ohm. This makes the numbers easy.)

Problem 7.



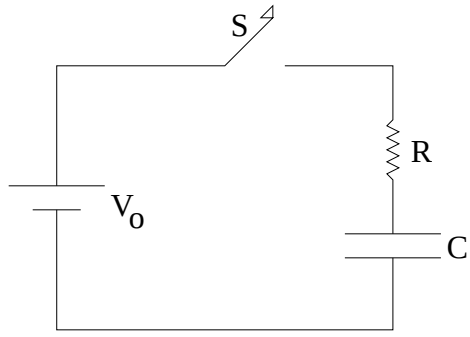
Find the currents I_1, I_2, and I_3 in the circuit above.

Problem 8.



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 9.



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 10.**

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.

Part III

Magnetostatics

Week 5: Moving Charges and Magnetic Force

(Est 2/13-2/18)

- A charge moving through space is observed to deflect according to the rule:

$$\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B}) \tag{5.1}$$

  which we use to *define* the magnetic field $\boldsymbol{B}$ much as we defined the electric field in terms of the force observed and described by Coulomb's Law.

  For the moment we will ignore just how vB got there, as we live in a locally uniform magnetic field due to the Earth all the time and can discover magnetic materials in nature so natural sources of magnetism are ubiquitous.

- This translates into:

$$\boldsymbol{F} = I(d\boldsymbol{\ell} \times \boldsymbol{B}) \tag{5.2}$$

  for a small (differential) segment of wire carrying a current I in a magnetic field vB. Magnetic fields exert forces on current carrying wires.

- Motion of a point charge in the plane perpendicular to a uniform magnetic field is therefore *circular*:

$$|\boldsymbol{F}| = qvB = \frac{mv^2}{r} \tag{5.3}$$

(Newton's second law plus definition of centripetal acceleration). It has an angular velocity given by:

$$\omega_{\text{cyclotron}} = \frac{qB}{m} \tag{5.4}$$

independent of its *speed*. This is called the *cyclotron frequency*.

- You should be able to derive/explain:
  - A velocity selector (region of crossed fields).
  - A cyclotron.
  - 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:

$$\boldsymbol{m} = NIA\hat{\boldsymbol{n}} \tag{5.5}$$

where N is the number of turns, I is the current, A is the area, and $\hat{\boldsymbol{n}}$ is the right-handed normal to the plane of the loop.

- The *torque* on a magnetic dipole in a uniform magnetic field is:

$$\boldsymbol{\tau} = \boldsymbol{m} \times \boldsymbol{B} \tag{5.6}$$

Associated with this are its potential energy:

$$U = -\boldsymbol{m} \cdot \boldsymbol{B} \tag{5.7}$$

and its force in a *non*-uniform magnetic field:

$$\boldsymbol{F} = -\boldsymbol{\nabla} U = \boldsymbol{\nabla}(\boldsymbol{m} \cdot \boldsymbol{B}) \tag{5.8}$$

Magnetic dipoles align with the field due to the torque, and then follow the field back to where it is stronger, just as do electric dipoles. Students have experienced this with toy magnets and refrigerator magnets from when they were very small – this is why bar magnets attract one another.

You should be able to compute the magnetic moment of simple current loops, although we'll get more practice at this in the next chapter/week.

5.1 Homework for Week 5

(Due 2/18/09)

Problem 1.

A particle with mass m and charge q a has a velocity $\boldsymbol{v}$ perpendicular to a uniform magnetic field $\boldsymbol{B}$ (with magnitude $B = |\boldsymbol{B}|$). 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 2.

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 V 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 on 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 3.

A nonconducting rod of total mass M and length L has a charge Q uniformly distributed along it. It is pivoted around one end and is rotating in the $x-y$ plane around the z-axis at angular frequency ω.

a) Consider a small bit of charge dq a distance r from the pivot and compute its average magnetic moment in the z-direction, dm_z.

b) Intgrate this result and find the total magnetic (dipole) moment of the rotating rod m_z

c) Show that the result can be expressed as $m_z = \frac{Q}{2M} L_z$ where L_z is the angular momentum of the rod about the pivot (that is to say, in the z-direction).

Problem 4.

Using the insight gained from the previous problem, show that the magnetic moment of a uniform nonconducting *disk* of charge Q, mass M, and radius R revolving at angular velocity about the z-axis is $m_z = \frac{Q}{2M} L_z$.

*** Problem 5.**

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.

Problem 6.

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 $\boldsymbol{B} = B_0 \hat{\boldsymbol{z}}$. Find the frequency ω_p with which the magnetic moment *precesses* around the magnetic field.

*** Problem 7.**

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:

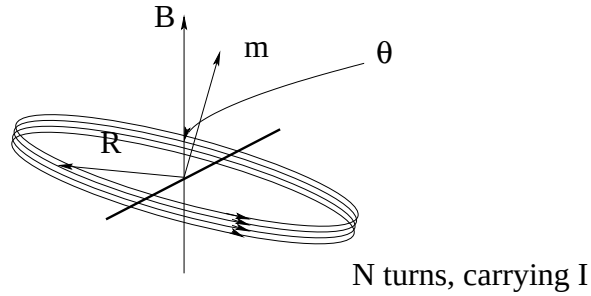
$$\boldsymbol{B} = \frac{k_m q_m \hat{\boldsymbol{r}}}{r^2}$$

at points "near" the end and outside of the solenoid itself. Note that $k_m = \mu_0/4\pi = 10^{-7}$ N-m/A^2 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 $\boldsymbol{r} = r\hat{\boldsymbol{r}}$ so its magnetic moment $\boldsymbol{m}$ is aligned with $\hat{\boldsymbol{r}}$. 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?

Problem 8.



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 $\boldsymbol{B} = B_0\hat{\boldsymbol{z}}$ 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.

Week 6: 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*:

$$\oint_S \boldsymbol{B} \cdot \hat{\boldsymbol{n}} dA = 4\pi k_m Q_{m,\text{in S}} = \mu_0 \int_{V/S} \rho_m dV = 0 \tag{6.1}$$

  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:

$$\boldsymbol{B} = k_m \frac{q\boldsymbol{v} \times \hat{\boldsymbol{r}}}{r^2} = \frac{\mu_0}{4\pi} \frac{q\boldsymbol{v} \times \hat{\boldsymbol{r}}}{r^2} \tag{6.2}$$

  where $\mu_0 = 4\pi \times 10^{-7}$ tesla-meter/ampere is called the *magnetic permeability of free space* and is the magnetic constant analoguous to ϵ_0, the dielectric permittivity of free space.

- If we consider a wire carrying a current $I = nqv_dA$ (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:

$$\begin{aligned} d\boldsymbol{B} &= k_m \frac{dq\boldsymbol{v}_d \times \hat{\boldsymbol{r}}}{r^2} \\ d\boldsymbol{B} &= k_m \frac{nqAd\ell\boldsymbol{v}_d \times \hat{\boldsymbol{r}}}{r^2} \\ d\boldsymbol{B} &= k_m \frac{nqv_dAd\boldsymbol{\ell} \times \hat{\boldsymbol{r}}}{r^2} \\ d\boldsymbol{B} &= k_m \frac{Id\boldsymbol{\ell} \times \hat{\boldsymbol{r}}}{r^2} \end{aligned}$$

  where $d\boldsymbol{\ell}$ is a differential length of the wire with a direction pointing *in the direction of the current*. This:

$$d\boldsymbol{B} = k_m \frac{Id\boldsymbol{\ell} \times \hat{\boldsymbol{r}}}{r^2} \tag{6.3}$$

  is known as the *Biot-Savart Law* for the magnetic field, and (one way or another) is the way most of the electrostatic fields we observe in nature come into being.

- The field of a long straight wire carrying a current I is:

$$\boldsymbol{B} = \frac{2k_mI}{r}\hat{\boldsymbol{\phi}} \tag{6.4}$$

  where $\hat{\boldsymbol{\phi}}$ 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*:

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{thru C}} = \mu_0 \int_{S/C} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA \tag{6.5}$$

  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:

$$\frac{dQ}{dt} = -\oint_S \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA \tag{6.6}$$

  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!

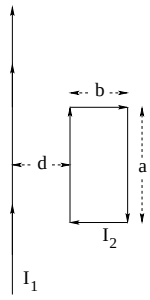
- Learn to use Ampere's Law to find the magnetic field of any cylindrically symmetric current distribution, a (long) solenoid, and a toroidal solenoid. (See homework)

- 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.

6.1 Homework for week 6

(Due 2/25/09)

Problem 1.



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 2.

Using Ampere's Law, find the magnetic field in all space produced by:

1. A solid conducting cylinder carrying a total current I.

2. Two cylindrical conductings shells carrying opposite currents (each equal to I in magnitude). The inner one has radius a, the outer one b.

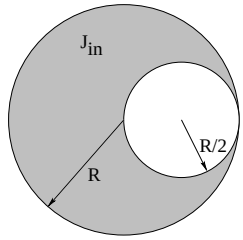
3. A solenoid with N turns and length L carrying current I in each turn (inside only, far from the ends).

4. A toroidal solenoid with N turns, inner radius a, outer radius b.

5. 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 3.



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* $\boldsymbol{J} = J\hat{\boldsymbol{z}}$ (and obviously $\boldsymbol{J} = 0$ in the hole). Find the magnetic field at all points inside the hole.

Problem 4.

Using the Biot-Savart law:

1. Find the $\boldsymbol{B}$-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).

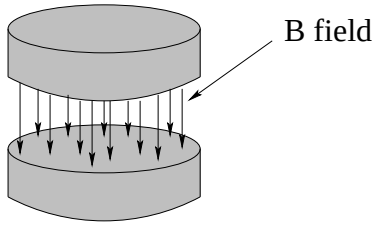
2. *Set up* the integral to be done to find the vB-field on the z axis of a *disk* in the $x - y$ plane of uniform charge density σ and radius a that is rotating with angular frequence ω around the z axis. **(A)** Do this integral (requires integration by parts a couple of times).

Problem 5.

Based on the *analogy* between electric and magnetic dipoles, deduce the probable form of the magnetic field of a spherical ball of charge Q, mass M, and radius R 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

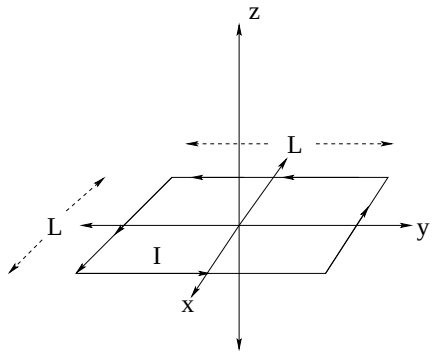
for any given multipolar moment, one merely has to compute that moment for a give charge-current density to discover the (far) field "for free".

Problem 6.



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.

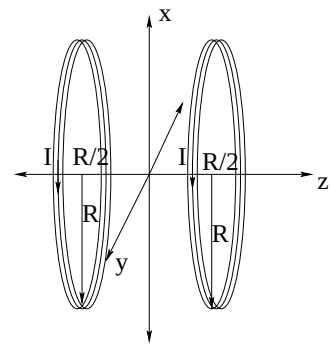
Problem 7.



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!

Problem 8.



(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.

Part IV

Electrodynamics

Week 7: 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*:

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{\ell} = -\frac{d}{dt} \int_{S/C} \boldsymbol{B} \cdot \hat{\boldsymbol{n}} dA \tag{7.1}$$

  The integral on the left is the *induced voltage* around the curve C.

- In this equation the minus sign is called *Lenz's Law* and tells us that the induced voltage decreases around the loop in the direction such that a flow of positive charge in that direction (the *induced current* if the loop is a conducting pathway) will *oppose the change* in the varying flux. If the flux is decreasing it will generate a magnetic moment that points in the direction that will increase it. If it is increasing it will generate a magnetic moment that points in the direction that will decrease it. This causes the *opposition* to motion noted in the motional voltage problems above.

- The flux through a conducting loop is directly proportional to the current through the loop itself or to the current through nearby sources of magnetic field that produce the flux. The constant of proportionality in either case depends solely on the *geometry* of the loop and source(s). That is, given a bunch of loops:

$$\phi_i = \sum_{j \neq i} M_{ij} I_j + L_i I_i \tag{7.2}$$

  where the M_{ij} are called the *mutual inductances* between the ith and jth loops and L_i is the *self inductance* of the ith loop.

- From this we can compute the *self*-induced (loop) voltages for simple current-carrying loops, in particular solenoids. To compute the self-inductance of a solenoid we begin with the result for the magnetic field inside an ideal solenoid from Ampere's Law:

$$B = \frac{\mu_0 N I}{L} \tag{7.3}$$

  (parallel to the solenoid axis). The current I creates a flux *per turn* that is equal to:

$$\phi_t = BA = \frac{\mu_0 N A I}{L} \tag{7.4}$$

  where A is the cross-sectional area of the solenoid. The total flux is thus:

$$\phi = NBA = \frac{\mu_0 N^2 A I}{L} = L_s I \tag{7.5}$$

  where L_s is the self-inductance of the solenoid. Clearly:

$$L_s = \frac{\mu_0 N^2 A}{L} \tag{7.6}$$

which depends *only on the geometry of the solenoid* just as the capacitance of an arrangement of conductors depended only on *their* geometry.

- The self-inductance of solenoids can be altered by wrapping them around suitable *magnetic materials* that enhance (para) or reduce (dia) the magnetic fields inside. Solenoids so constructed are ubiquitous in circuit design, where they are known as *inductors*; they are labelled with their inductance L in *Henries*, the SI unit of inductance:

$$1 \text{ Henry } = \frac{1 \text{ Volt} - \text{Second}}{\text{Ampere}} = 1 \text{ Ohm} - \text{Second} \tag{7.7}$$

- In terms of inductance:

$$V_L = -L\frac{dI}{dt} \tag{7.8}$$

is a statement of the voltage across an inductor using Faraday's Law.

- Mutual inductance is the basis of a number of devices, in particular a center-tap full-wave rectifier commonly used in e.g. DC power supplies or AM radios and in *transformers*, an essential component of the power distribution grid. If one imagines *two* solenoids, one with N_1 turns and cross sectional area A and a second one with N_2 turns wrapped *around* the first (so all of the flux (per turn) in the first passes through the loops of the second:

$$\phi_t = \frac{\mu_0 N_1 A I_1}{L} \tag{7.9}$$

for the first solenoid, so:

$$\phi_2 = N_2\frac{\mu_0 N_1 A I_1}{L} \tag{7.10}$$

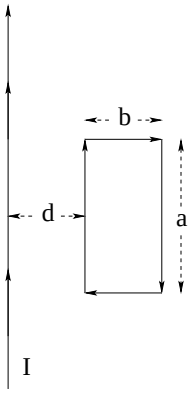
is the total flux through the second solenoid due to the current in the first. Thus:

$$M_{21} = \frac{\phi_2}{I_1} = \frac{\mu_0 N_1 N_2 A}{L} = M_{12} = M \tag{7.11}$$

7.1 Homework for week 7

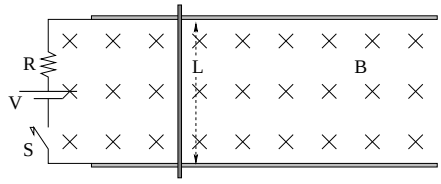
(Due 3/11/09)

Problem 1.



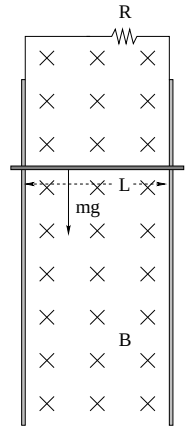
A switch is closed and a long straight wire builds up a current $I(t) = I_0(1 - e^{-\frac{t}{\tau}})$. A rectangular loop of wire with resistance R and dimensions $a \times b$ is a distance d away as shown. Find: a) the flux through the loop due to the wire; b) the induced voltage in the loop; c) the induced current in the loop; d) the force between the loop and the wire (remember homework problem 6.1).

Problem 2.



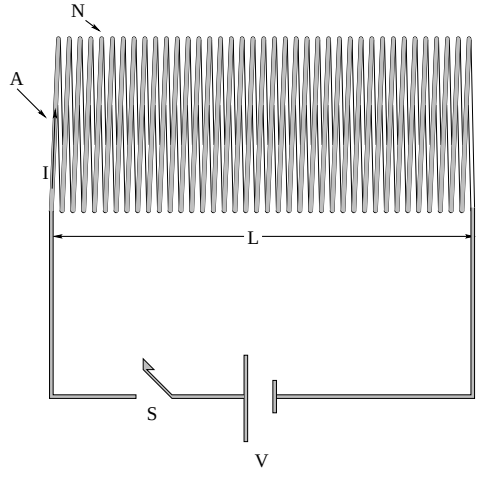
A rod of length L and mass m sits at rest on two frictionless conducting rails that sit in a plane perpendicular to a magnetic field as shown. At time $t = 0$ a switch S is closed connecting a voltage V that goes through a resistance R and the rod. The rod begins to move from a $x = 0$. Find: a) the current in the loop as a function of time; b) the velocity of the rod as a function of time.

Problem 3.



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 current in the loop as a function of time; b) the velocity of the rod as a function of time.

Problem 4.

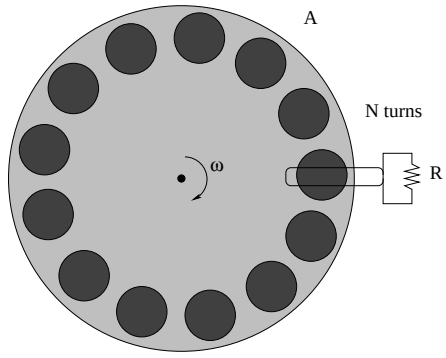


Find the self-inductance of the solenoid above that has N turns, length L, resistance R, and cross sectional are A. Then find the current $I(t)$ in the circuit assuming that the switch S is closed at time $t = 0$.

Problem 5.

A toroidal solenoid with a square cross section, that has inner radius R and square side $a \ll R$ and N turns. Begin with Ampere's Law to find the field given I, compute the flux, and from that find the self-inductance.

Problem 6.

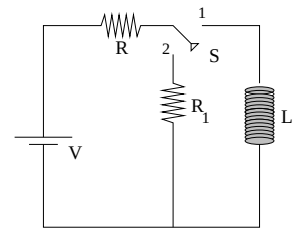


A magnetic braking system is drawn above. A wheel has M powerful permanent magnets mounted around the rim. Each magnet produces a uniform field B across a cross-sectional area A. As the wheel spins at angular velocity ω, the magnets cross in front of a coil with N turns in a circuit with a resistance 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.
- Compute the *power* during the ramp up and ramp down phases.
- Using this power as the average power, compute the total energy dissipated as heat in the resistor as a function of ω.

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 7.



1. In the circuit above, switch S is closed in position 1 at time $t = 0$. Using Kirchhoff's voltage rule, find (derive) and solve the differential equation for $I(t)$, the current in the circuit loop. Plot this function "generically" in units of τ (the exponential time constant for this circuit), a few τ out from $t = 0$. What is τ?

2. At time $t = \tau$, the switch is *quickly* moved from position 1 to position 2 (so fast that the current is uninterrupted during the transition). As before, derive $I(t)$ and plot it a few τ out. (Hint: You can "restart the clock" in terms of $t' = t - \tau$, right? And you can start at $t' = 0$ at current $I_0 = I(\tau)$ from the first part. That makes this problem relatively simple.)

Week 8: Alternative Current Circuits

(Est 3/4-3/18)

- **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:

$$\begin{aligned} \phi_m &= \boldsymbol{B} \cdot NA\hat{\boldsymbol{n}} = NBA\cos(\omega t) && (8.1) \\ V(t) &= -\frac{d\phi_m}{dt} = NBA\omega\sin(\omega t) && (8.2) \end{aligned}$$

  We will from now on treat "arbitrary" harmonic alternating voltage sources as having the form:

$$V(t) = V_0\sin(\omega t) \tag{8.3}$$

  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:

$$\begin{aligned} V &= 120\sin(\omega t) + 120\sin(\omega t \pm 2\pi/3) \\ &= 240\sin(\pi/3)\sin(\omega t \pm \pi/3) \\ &= 208\sin(\omega t \pm \pi/3) && (8.4) \end{aligned}$$

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 8.1: 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.

and 240 is similarly the difference between two 120 volt lines that are completely out of phase. Do *not* use this table as an authoritative guide to electrical main supplies around the world; there are many such authoritative guides and tables available on the internet [1] .

It is worth mentioning that (unfortunately) 60 Hz is a *particularly unfortunate* choice for distribution frequency because it is in "resonance" with certain cardiac frequencies and hence unusually likely to defibrillate the human heart. As little as 10 mA of 60 Hz AC across the heart can kill a person. It requires roughly five times as much DC (50 mA) to be equivalently dangerous!

- The reason for using such low frequencies is that AC does not flow uniformly through a conductor – it is 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-

[1] Wikipedia: http://www.wikipedia.org/wiki/Mains_electricity. See also the many links in this article.

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:

$$V_s = N_s \frac{d\phi_m}{dt} \tag{8.5}$$

$$V_p = N_p \frac{d\phi_m}{dt} \tag{8.6}$$

or (taking the ratios of these two equations, in order)

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \tag{8.7}$$

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 $I^2R \approx 1$ Watt used in the *supply line.*

  However, you then turn on the *rest* of your lights, your refrigerator kicks on, your AC starts up. Your house is now drawing more like 100 Amperes (delivered in parallel to the many appliances) and is using order of 10000 Watts. *So is the supply line!* Half of the energy being delivered to your home is wasted as heat along the way. A second consequence is that the *voltage* at your house is reduced to a fraction of the nominal voltage as you turn on more appliances and more of the voltage drop occurs across the supply resistance!

  The solution is to *transmit at high voltage and low current* and *use at low voltage and high current.* If we step up the voltage by (say) 10,000 Volts (real long distance transmission is at much higher voltages than this) then in order to deliver the same *power* at the far end, instead of delivering 100 Amps at 100 volts one can deliver 1 Amp at 10,000

Volts! The resistive heating of the supply line is back to 1 Watt out of 10,000 delivered. Here the square in I^2R becomes your *friend* – delivering 10 kW at 100,000 V requires only 0.1 A and uses only 0.01 W heating the wire.

This is good for transmission, but bad for utilization. 100,000 volts can arc an appreciable distance through even *dry* air; that's why the insulators on high voltage transmission towers are so long! We'd hate to get electrocuted every time we changed a light bulb as power arced out of the socket through our bodies on the way to ground. With an entire power plant delivering the energy, even the (mere) 16,000 volt lines that run down the streets can literally make your body explode if you should stray within a few cm of a supply line. 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 $\sim$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:

$$\frac{Q}{C} - L\frac{dI}{dt} = 0 \tag{8.8}$$

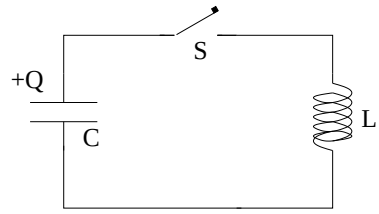


Figure 8.1: Undriven LC circuit

where

$$I = -\frac{dQ}{dt} \tag{8.9}$$

If we substitute this relation in for the I's and divide by L, we get the following second order, linear, homogeneous ordinary differential equation:

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = 0 \tag{8.10}$$

We recognize this as the differential equation for a *harmonic oscillator!* To solve it, we "guess"[2]:

$$Q(t) = Q_0 e^{\alpha t} \tag{8.11}$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{1}{LC} = 0 \tag{8.12}$$

We solve for:

$$\alpha = \pm i\sqrt{\frac{1}{LC}} = \pm i\omega_0 \tag{8.13}$$

and get:

$$Q(t) = Q_{0+}e^{+i\omega_0 t} + Q_{0-}e^{-i\omega_0 t} \tag{8.14}$$

or (taking the real part and using the initial conditions):

$$Q(t) = Q_0 \cos(\omega_0 t) \tag{8.15}$$

- **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

[2]Not really.

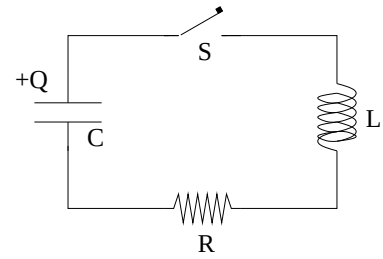


Figure 8.2: Undriven LRC circuit

closed and current begins to flow. If we apply Kirchhoff's voltage/loop rule to the circuit, we get:

$$\frac{Q}{C} - L\frac{dI}{dt} - IR = 0 \tag{8.16}$$

where

$$I = -\frac{dQ}{dt} \tag{8.17}$$

If we substitute this relation in for the I's and divide by L, we get the following second order, linear, homogeneous ordinary differential equation:

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{Q}{LC} = 0 \tag{8.18}$$

We recognize this as the differential equation for a *damped harmonic oscillator*. To solve it, we "guess"[3]:

$$Q(t) = Q_0 e^{\alpha t} \tag{8.19}$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{R}{L}\alpha + \frac{1}{LC} = 0 \tag{8.20}$$

We solve for:

$$\begin{aligned} \alpha &= -\frac{R}{2L} \pm \frac{\sqrt{\left(\frac{R}{L}\right)^2 - \frac{4}{LC}}}{2} \\ &= -\frac{R}{2L} \pm i\omega_0\sqrt{1 - \frac{R^2C}{4L}} \end{aligned}$$

[3]Not really.

$$\begin{aligned} &= -\frac{R}{2L} \pm i\omega_0 \sqrt{1 - \frac{\tau_L}{4\tau_R}} \\ &= -\frac{R}{2L} \pm i\omega' \end{aligned} \tag{8.21}$$

where $\tau_L = R/L$ $\tau_C = 1/RC$, $\omega' =_0 \sqrt{1 - \frac{\tau_L}{4\tau_R}}$, and our final solution looks like:

$$Q(t) = Q_0 e^{\frac{Rt}{2L}} \cos(\omega' t) \tag{8.22}$$

(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 $t \to \infty$.

- **Resistance R across an AC voltage:**

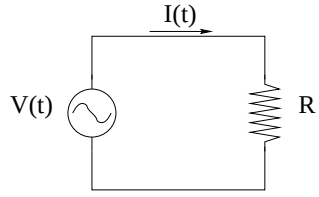


Figure 8.3: AC voltage across R

We use Kirchhoff's voltage rule and Ohm's Law to get:

$$V_0 \sin(\omega t) - IR = 0 \tag{8.23}$$

or

$$I_R(t) = \frac{V_0}{R} \sin(\omega t) \tag{8.24}$$

and we see that the current is *in phase* with the voltage drop across a resistor.

- **Capacitance C across an AC voltage:**

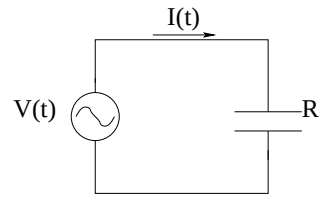


Figure 8.4: AC voltage across CR

We use Kirchhoff's voltage rule and the definition of capacitance to get:

$$V_0 \sin(\omega t) - \frac{Q}{C} = 0 \tag{8.25}$$

We can solve for $Q(t)$:

$$Q(t) = CV_0 \sin(\omega t) \tag{8.26}$$

Finally, we note that:

$$\begin{aligned} I_C(t) &= \frac{dQ(t)}{dt} = (\omega C)V_0 \cos(\omega t) \\ &= (\omega C)V_0 \sin(\omega t + \pi/2) = I_0 \sin(\omega t + \pi/2) \end{aligned} \tag{8.27}$$

where

$$I_0 = (\omega C)V_0 = \frac{V_0}{\chi_C} \tag{8.28}$$

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 *capacitative reactance*, the "resistance" of a capacitor to alternating voltages.

- **Inductance L across an AC voltage:**

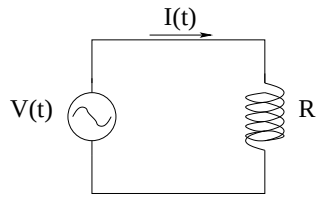


Figure 8.5: AC voltage across L

We use Kirchhoff's voltage rule and the definition of capacitance to get:

$$V_0 \sin(\omega t) - L\frac{dI}{dt} = 0 \tag{8.29}$$

We can solve for $dI(t)$:

$$dI = \frac{V_0}{L}\sin(\omega t)dt \tag{8.30}$$

We integrate both sides to get:

$$\begin{aligned} I_L(t) &= \int \frac{V_0}{L}\sin(\omega t)dt \\ &= \int \frac{V_0}{\omega L}\sin(\omega t)\ \omega dt \\ &= \frac{V_0}{\omega L}\cos(\omega t) && (8.31) \\ &= \frac{V_0}{\omega L}\sin(\omega t - \pi/2) && (8.32) \\ &= I_0 \sin(\omega t - \pi/2) && (8.33) \\ & && (8.34) \end{aligned}$$

where

$$I_0 = \frac{V_0}{\omega L} = \frac{V_0}{\chi_L} \tag{8.35}$$

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

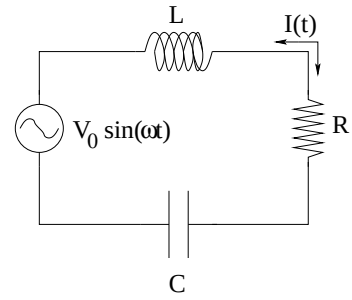


Figure 8.6: A LRC (tank) circuit.

to this circuit and get:

$$V_0 \sin(\omega t) - L\frac{dI}{dt} - RI - \frac{Q}{C} = 0 \tag{8.36}$$

or

$$V_L + V_R + V_C = V_0 \sin(\omega t) \tag{8.37}$$

or

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{1}{LC}Q = \frac{V_0}{L}\sin(\omega t) \tag{8.38}$$

There are a number of way 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 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:

$$Q(t) = Q_p(t) + Q_h(t) \tag{8.39}$$

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*:

$$I(t) = I_0 \sin(\omega t - \phi) \tag{8.40}$$

then solving the problem is easy[4]. We begin by noting the voltage drops across all three circuit elements in terms of $I(t)$:

$$\begin{aligned} V_R &= I_0 R \sin(\omega t - \phi) && (8.41)\\ V_L &= I_0 \chi_L \sin(\omega t - \phi + \pi/2) && (8.42)\\ V_C &= I_0 \chi_C \sin(\omega t - \phi - \pi/2) && (8.43) \end{aligned}$$

or

$$\begin{aligned} I_0 R \sin(\omega t - \phi) &+ I_0 \chi_L \sin(\omega t - \phi + \pi/2) \\ &+ I_0 \chi_C \sin(\omega t - \phi - \pi/2) = V_0 \sin(\omega t) \end{aligned} \tag{8.44}$$

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 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.

[4]This 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.

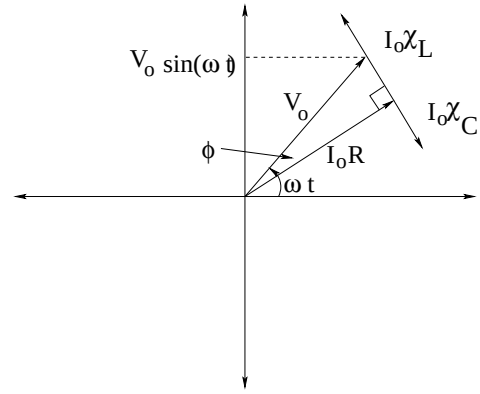


Figure 8.7: A phasor diagram for the *LRC* circuit.

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:

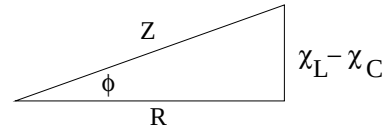


Figure 8.8: The impedance diagram for the *LRC* circuit.

$$Z = \sqrt{R^2 + (\chi_L - \chi_C)^2} \tag{8.45}$$

so that

$$I_0 = \frac{V_0}{Z} \tag{8.46}$$

and

$$\phi = \tan^{-1}\left(\frac{\chi_L - \chi_C}{R}\right) \tag{8.47}$$

- **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:

$$I_R = \frac{V_0}{R}\sin(\omega t) \tag{8.48}$$

$$I_L = \frac{V_0}{\chi_L}\sin(\omega t - \pi/2) \tag{8.49}$$

$$I_C = \frac{V_0}{\chi_C}\sin(\omega t + \pi/2) \tag{8.50}$$

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:

$$\begin{aligned} I_{\rm tot} &= \frac{V_0}{R}\sin(\omega t) + \frac{V_0}{\chi_L}\sin(\omega t - \pi/2) + \frac{V_0}{\chi_C}\sin(\omega t + \pi/2) \\ &= \frac{V_0}{Z}\sin(\omega t - \phi) \\ &= I_0 \sin(\omega t - \phi) \end{aligned} \tag{8.51}$$

In this expression, a bit of contemplation should convince you that the impedance Z for this circuit is given by the entirely reasonable:

$$\frac{1}{Z} = \sqrt{\left(\frac{1}{R^2} + (\frac{1}{\chi_C} - \frac{1}{\chi_L})^2\right)} \tag{8.52}$$

which we recognize as the phasor equivalent of the familiar rule for reciprocal addition of resistances in parallel, and:

$$\begin{aligned} \phi &= \tan^{-1}\left(\frac{\frac{1}{\chi_C} - \frac{1}{\chi_L}}{\frac{1}{R}}\right) \\ &= \tan^{-1}\left(\frac{RC(\omega^2 - \omega_0^2)}{\omega}\right) \end{aligned} \tag{8.53}$$

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 $frac1Z$ 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:

$$P_R = \frac{V_0^2 \sin^2(\omega t)}{R} \tag{8.54}$$

so that the average power delivered to the circuit is:

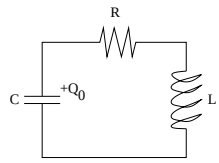
$$< P >=< P_R >= \frac{V_0^2}{2R} \tag{8.55}$$

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.

8.1 Homework for week 8

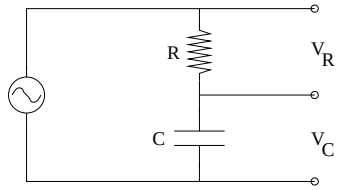
(Due 3/18/09)

Problem 1.



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 2.



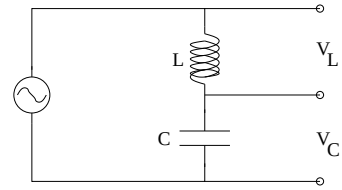
In the circuit above, the AC voltage is $V_0 \cos(\omega t)$. Find:

1. 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.
2. The voltage $V_R(t)$ across the resistor. Factor your answer out so that it is in terms of the dimensionless ωRC.
3. 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.

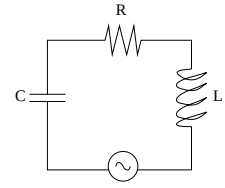
Problem 3.



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 LR versus the LC filters, respectively? Think about this: Not all loads are *resistive*...

Problem 4.



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.

Problem 5.

Second, *derive* the expression:

$$P_{\rm av}(\omega) = I_{\rm av}^2 R = \frac{V_{\rm rms}^2 R\omega^2}{L^2(\omega^2 - \omega_0^2) + \omega^2 R^2}$$

(noting carefully and proving along the way that the average power delivered to the inductor and capacitor is zero). In this expression $\omega_0 = 1/LC$ as you should fully understand at this point.

Then show that for a sharply peaked resonance (one with large Q)

$$\Delta\omega \approx \frac{R}{L}$$

so that

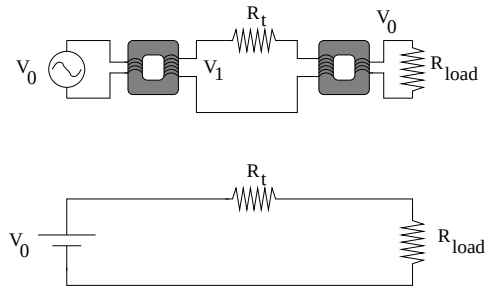
$$Q = \frac{\omega_0}{\Delta\omega} \approx \frac{\omega_0 L}{R}$$

where $\Delta\omega$ is the full width at half maximum of the power curve you derive in the first part.

You may find the following factorization useful:

$$\omega^2 - \omega_0^2 = (\omega - \omega_0)(\omega + \omega_0)$$

Problem 6.



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 alterntives 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 $R_{\rm load}$ 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 $P_{\rm load}$ (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 $P_t \ll P_{\rm load}$. Find a relationship between the ratios:

$$\frac{V_0}{V_1}$$

and

$$\frac{P_t}{P_{\rm load}}$$

that proves that Tesla's solution wins (and by how much it wins, given "reasonable" estimates for $R_t/R_{\rm load}$).

Week 9: Maxwell's Equations and Light

- 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*:

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{\ell} = \mu_0 \left(\int_{S/C} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} dA + \frac{d}{dt} \epsilon_0 \int_{S/C} \boldsymbol{E} \cdot \hat{\boldsymbol{n}} dA \right)$$

  Both of these latter two integrals must be evaluated with the *same* surface S, but given this they sum together to give the same invariant current for *all* the surfaces S that are bounded by the closed curve C.

- In this new, *correct* version of Ampere's Law, you can see Maxwell's contribution: the *Maxwell Displacement Current* produced by a *time varying electric field*:

$$I_{MDC} = \frac{d}{dt} \epsilon_0 \int_{S/C} \boldsymbol{E} \cdot \hat{\boldsymbol{n}} dA$$

- It is worth writing down the complete set of trading cards, suitable for engraving:

$$\oint_S \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA = \frac{1}{\epsilon_0}\int_{V/S}\rho_e dV \tag{9.1}$$

$$\oint_S \boldsymbol{B}\cdot\hat{\boldsymbol{n}}dA = \mu_0\int_{V/S}\rho_m dV = 0 \tag{9.2}$$

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{\ell} = \mu_0\left(\int_{S/C}\boldsymbol{J}\cdot\hat{\boldsymbol{n}}dA + \frac{d}{dt}\epsilon_0\int_{S/C}\boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA\right) \tag{9.3}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{\ell} = -\frac{d}{dt}\int_{S/C}\boldsymbol{B}\cdot\hat{\boldsymbol{n}}dA \tag{9.4}$$

- Physicists usually rearrange them to make the equations connecting fields to *sources* stand out from the equations that have no source terms (because we have yet to see a magnetic monopole):

$$\oint_S \boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA = \frac{1}{\epsilon_0}\int_{V/S}\rho_e dV \tag{9.5}$$

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{\ell} - \frac{d}{dt}\mu_0\epsilon_0\int_{S/C}\boldsymbol{E}\cdot\hat{\boldsymbol{n}}dA = \mu_0\int_{S/C}\boldsymbol{J}\cdot\hat{\boldsymbol{n}}dA \tag{9.6}$$

$$\oint_S \boldsymbol{B}\cdot\hat{\boldsymbol{n}}dA = 0 \tag{9.7}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{\ell} + \frac{d}{dt}\int_{S/C}\boldsymbol{B}\cdot\hat{\boldsymbol{n}}dA = 0 \tag{9.8}$$

  This way, the symmetry *is compelling!* Two inhomogeneous equations have source terms connected to electric charge, two homogeneous equations have the *same form* but lack the source terms, at least until monopoles are discovered.

- If one applies these equations to a *source-free volume of space* where electric and magnetic fields are varying, one can show that they lead to the following *wave equations* for the *electromagnetic field* propagating in (say) the z-direction:

$$\frac{\partial^2\boldsymbol{E}}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} = 0 \tag{9.9}$$

$$\frac{\partial^2\boldsymbol{B}}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{B}}{\partial t^2} = 0 \tag{9.10}$$

  The $\frac{\partial^2}{\partial z^2}$ symbol in this expression, let me remind you, just means to take the derivative of the functions $\boldsymbol{E}(\boldsymbol{x},t)$ and $\boldsymbol{B}(\boldsymbol{x},t)$ with respect

to the z-coordinate only, pretending that the other coordinates are constants. In this equation,

$$c = \sqrt{\frac{k_e}{k_m}} = \frac{1}{\sqrt{\epsilon_0 \mu_0}} = 3 \times 10^8 \text{meters per second} \tag{9.11}$$

is the *speed of light in a vacuum*, which we can see is *completely determined* from Maxwell's equations.

Since Maxwell's equations are laws of nature and expected to hold in all inertial reference frames, it is entirely *reasonable* to expect the speed of light to be constant in all reference frames! This postulate, together with some very simple assumptions about coordinate transformations, suffices to derive the theory of relativity!

- We will study the details of at least certain simple solutions to these wave equations over the next few weeks. For the moment, the most important solution for you to learn is:

$$E_x(z,t) = E_{0x} \sin(kz - \omega t) \tag{9.12}$$
$$B_y(z,t) = B_{0y} \sin(kz - \omega t) \tag{9.13}$$

known as a *harmonic plane wave* travelling in the z-direction. Note that E_x and B_y are *in phase* and do not have independent amplitudes – their amplitudes are connected by Maxwell's equations (Faraday or Ampere's law) and $E_x = cB_y$. There is an identical pair of solutions with a different *polarization*:

$$E_y(z,t) = E_{0y} \sin(kz - \omega t) \tag{9.14}$$
$$B_x(z,t) = -B_{0x} \sin(kz - \omega t) \tag{9.15}$$

that also propagate in the z-direction, as determined from the derivation of the wave equations above.

In these equations, note well that:

$$k = \frac{2\pi}{\lambda} \tag{9.16}$$

is the *wave number* of the wave, where λ is the *wavelength* of the harmonic wave, while:

$$\omega = \frac{2\pi}{T} \tag{9.17}$$

is the *angular frequency* of the wave. The wavelength is thus the "spatial period" of the wave, where T is the "temporal period" of the wave that harmonically oscillates in space and time. This wave propagates in the *positive* z-direction as can be seen by considering $kz - \omega t = k(z - \frac{\omega}{k}t) = k(z - ct)$. Note well that this uses the result that:

$$c = \frac{\lambda}{T} = \frac{\omega}{k} \tag{9.18}$$

for a harmonic wave.

- The flow of energy in an electromagnetic wave (and field in general) can be determined from the *Poynting vector*:

$$\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B}) \tag{9.19}$$

The magnitude of the Poynting vector is called the *intensity* of the electromagnetic wave – the energy per unit area per unit time or power per unit area being transported by the wave in the direction of its motion:

$$I = \frac{dP}{dA} = \frac{d}{dA}\frac{dU}{dt} = |S| \tag{9.20}$$

where U is the energy in the wave. To speak more mathematically precisely to communicate the transport of *power* (energy per unit time, in watts) across some given surface A, one evaluates the *flux of the Poynting vector through the surface*:

$$P_A = \int_A \boldsymbol{S} \cdot \hat{\boldsymbol{n}}\, dA \tag{9.21}$$

As you can see one just cannot get away from flux integrals as a way of representing the "flow" of energy, current, fluid, or $\boldsymbol{E}$ or $\boldsymbol{B}$ field through a surface! As such, it is a very important idea to conceptually master.

- The Poynting vector can be understood and *almost* derived by adding up the total energy in the electric and magnetic fields in a volume of space being transported perpendicular to a surface A. In a time Δt, all of the energy in a volume $\Delta V = A\ c\Delta t$ goes through the surface at the end. This is:

$$\Delta U = (\frac{1}{2}\epsilon_0 E_x^2 + \frac{1}{2\mu_0}B_y^2)A\ c\Delta t \tag{9.22}$$

If we use $|E_x| = c|B_y|$ (see above) for a wave travelling in the z-direction and do a bit of algebra, we can see that:

$$\frac{\Delta U}{A\Delta t} = \frac{1}{\mu_0}|\boldsymbol{E}_x||\boldsymbol{B}_y| \tag{9.23}$$

which is just the Poynting vector magnitude in the z-direction for these two field components.

- The electromagnetic field also carries *momentum*, solving the dilemma of the "missing momentum" left over from our consideration of the magnetic force and the failure of Newton's third law. The field momentum is rather difficult to derive in a *simple* way, but it can *somewhat* be understood by assuming that the field *electrically* polarizes atoms that it sweeps over in such a way that it exerts a *magnetic* force along the direction of motion of the electromagnetic wave. We'll explore this with a problem later. The momentum density of the electromagnetic field is:

$$|p_f| = \frac{U}{c} \tag{9.24}$$

and we can consider the net momentum transported per unit area per unit time by the electromagnetic field perpendicular to a surface A to be:

$$P_r = \frac{I_{\text{thru A}}}{c} \tag{9.25}$$

This quantity is called the *radiation pressure* and it is partially responsible for the *solar wind*, created as sunlight pushes gas molecules away from the sun. Light "sails" have also been proposed as a propulsion for getting around inside the solar system without rocket fuel. We will explore both of these ideas with homework problems.

To use radiation pressure properly, one has to compute the force it exerts on a surface. This force will depend on certain things, such as whether or not the radiation is perfectly absorbed or perfectly reflected and (eventually) the relative velocity of source and target (as the incident and reflected waves can be doppler shifted, affecting the momentum transfer). In the simplest cases (perfect absorption or reflection) the force is best computed by using an expression such as:

$$F_S = \frac{1}{c}\int_A \boldsymbol{S}\cdot\hat{\boldsymbol{n}}\; dA \tag{9.26}$$

that is, the flux of the Poynting vector yields the power transferred to a (perfectly absorbing) surface, and $1/c$ of the power is the effective force exerted along the line of the original Poynting vector. If the radiation is reflected, one has to construct a such quantity evaluated (with the same power) with respect to the direction of the angle of reflection, and vector sum the forces. In the simplest case of normal absorption or reflection:

$$F_S = \frac{SA}{c} \tag{9.27}$$

or

$$F_S = \frac{2SA}{c} \tag{9.28}$$

respectively.

- Electromagnetic radiation is produced when electrical charges *accelerate* (this follows from construction the *inhomogeneous wave equations* for the electromagnetic fields directly from Maxwell's equations, where moving charge and current terms become the sources of the time varying fields). This has two important consequences you should be aware of.

  The first is that *oscillating electrical dipoles* (a very reasonable model for any atom or molecule that has been "kicked" one way or another) act as *antennae* and radiate away electromagnetic radiation. 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:

  $$I(\theta) = \frac{I_0}{r^2}\sin(\theta) \tag{9.29}$$

  (and is azimuthally symmetric about the z-axis). 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.

  The second is that it becomes impossible to build a simple classical model for a bound atom that looks, as one might reasonably expect, like a heavy positive nucleus being orbited by negative light electrons in "planetary" orbits. Or any other e.g. harmonic oscillator model for bound atoms. Atoms (and ordinary matter) becomes *unstable* under any classical physical model, radiating away all of of their energy in

an extraordinarily short period of time in model calculations and collapsing. Maxwell's equations and Newton's Laws are *inconsistent* and cannot both be correct. Experimentally, it has been determined that Maxwell's equations are indeed correct (although they are part of a bigger theory we are still working on) but *Newton's Laws are not!*

The discovery that accelerated charges radiate electromagnetic energy was thus the death knell of classical physics! It took some twenty or thirty years, but physicists invented the theory of *quantum mechanics* to explain a wide variety of otherwise puzzing phenomena and resolve the conflict between Maxwell's equations and classical dynamics. Classical physics (Newton's laws) hold in a *particular limit* of quantum theory but break down at small (atomic) length and time scales where Planck's Constant h becomes important.

9.1 Homework for week 9

Problem 1.

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 $\boldsymbol{E}$ and $\boldsymbol{B}$ are independent of x and y) that:

$$\frac{\partial E_x}{\partial z} = \frac{\partial B_y}{\partial t} \tag{9.30}$$

$$\frac{\partial B_y}{\partial z} = \mu_0\epsilon_0\frac{\partial E_x}{\partial t} \tag{9.31}$$

and

$$\frac{\partial E_y}{\partial z} = \frac{\partial B_x}{\partial t} \tag{9.32}$$

$$\frac{\partial B_x}{\partial z} = \mu_0\epsilon_0\frac{\partial E_y}{\partial t} \tag{9.33}$$

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 2.

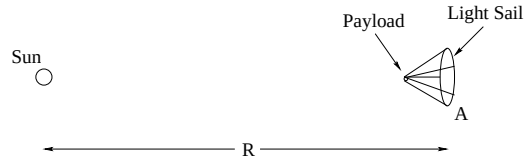
Show that $f(z \pm vt)$ satisfies the wave equation:

$$\frac{\partial^2 f}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{9.34}$$

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 3.



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 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 4.

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.

1. Find the net resistance R of the resistor.
2. Find the current I through the resistor.
3. Find the electric field inside the resistive material.
4. Find the magnetic field as a function of distance from the cylinder axis inside the resistive material (assume that its permeability is μ_0).
5. Evaluate the *Poynting vector* $\boldsymbol{S}$ at an arbitrary point on the *cylindrical surface* of the resistor.
6. 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 5.

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 accelarated 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 6.

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 $I(t) = I_0(1-e^{-t/\tau})$.

1. Find $B_z(t)$ inside the solenoid.

2. Find the induced electrical field at an arbitrary point inside the solenoid (say, at a distance r from its axis).

3. Find the magnitude and direction of the Poynting vector on an imagined surface of constant radius just inside the windings at radius R.

4. Compute the flux of the Poynting vector into the volume of the solenoid.

5. 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 7.

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 radation 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 8.

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 solenoid. Show that the result equals $P_C = V_C I$, the power being delivered to the solenoid. (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.)

Week 10: Light

- The speed of light in a medium is:

$$v_{\text{medium}} = \frac{c}{n} \tag{10.1}$$

  n is called the *index of refraction* of the medium. You need to know the following *approximate* indices of refraction to work problems: Air: $n_a \approx 1$. Water: $n_w \approx 4/3$. Glass: $n_g \approx 3/2$. Any others needed will be given in the problem in context.

- The index of refraction is not constant – it varies with the *frequency* of the light: $n(\omega)$, a phenomena known as *dispersion*.

- The Law of Reflection:

  The angle of incidence equals the angle of reflection,

$$\theta_i = \theta_\ell \tag{10.2}$$

- Snell's Law:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.3}$$

- Fermat's Principle:

  Light takes the path that minimizes the time of flight between any two points. Both the law of reflection and Snell's law can be derived from Fermat's principle.

- Critical Angle, Total Internal Reflection:

  Light passing from a dense medium n_2 to a less dense medium $n_1 < n_2$ is *totally internally reflected* if the angle of incidence is greater than:

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right) \tag{10.4}$$

- Polarization:

  We describe the orientation and phase of the two components of the *electric* field component for a given fixed harmonic frequency as the *polarization* of the harmonic wave.

- Unpolarized Light:

  Unpolarized light is light for which the polarization vector is constantly shifting its direction around. On average, unpolarized light has its energy/intensity equally distributed between the two independent directions of polarization.

- Linear Polarization:

  Linear polarization occurs whenever the electric field vector oscillates consistently in a single vector direction in the plane perpendicular to propagation.

- Circularly Polarized Light:

  Circularly polarized light has the same electric field magnitude in the two independent polarization directions but the waves in these directions are $\pi/2$ out of phase:

$$\begin{aligned} \boldsymbol{E}(z,t) &= \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz-\omega t\pm\pi/2)+\frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz-\omega t) \\ \boldsymbol{E}(z,t) &= \frac{\sqrt{2}}{2}E_0\left(\pm\hat{\boldsymbol{x}}\cos(kz-\omega t)+\hat{\boldsymbol{y}}\sin(kz-\omega t)\right) \end{aligned} \tag{10.5}$$

  There are two independent *helicities* of circularly polarized light: right (clockwise/+) and left (anticlockwise/-) when facing *in* the direction of propagation).

- Elliptically Polarized Light:

  If the amplitudes of the two waves are (potentially) different *and* the two waves are (potentially) out of phase, the most general polarization state is that of *elliptical* polarization:

$$\boldsymbol{E}(z,t)=E_{0x}\hat{\boldsymbol{x}}\sin(kz-\omega t+\delta_x)+E_{0y}\hat{\boldsymbol{y}}\sin(kz-\omega t+\delta_y) \tag{10.6}$$

  In this expression, E_{0x} and E_{0y} may or may not be equal, and the phases δ_x and δ_y may or may not be zero *or* equal.

- Polarization by Absorption (Malus's Law):

  For an ideal polaroid filter that is otherwise fully transparent:

  $$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \tag{10.7}$$

  The transmitted light is fully linearly polarized in the direction of the **transmission axis** of the filter.

  If the light that is incident on the filter is already polarized, then only the *component* of the electric field vector that is *parallel* to the transmission axis is transmitted:

  $$E_{\text{transmitted}} = \boldsymbol{E} \cdot \hat{\boldsymbol{t}} = E_{\text{incident}} \cos(\theta) \tag{10.8}$$

  where θ is the angle between the direction of linear polarization of the incident light and a unit vector along the transmission axis. This implies that the transmitted intensity is given by:

  $$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \tag{10.9}$$

  This result is known as **Malus's law**.

- Polarization by Scattering:

  Rays scattered more or less at right angles to an atom, molecule, or speck of dust are linearly polarized **perpendicular to the plane of scattering.**

- Polarization by Reflection:

  Light that is reflected at a non-normal angle from a dielectric surface is (partially or completely) polarized **parallel to the surface**, which is also **perpendicular to the plane of reflection**. Light transmitted into the new medium is partially polarized the opposite way (by subtraction).

  The reflected light is *completely* polarized when the light is incident at the *Brewster angle*, where the reflected and refracted rays are perpendicular to each other, given by:

  $$\tan(\theta_b) = \frac{n_2}{n_1} \tag{10.10}$$

- Polaroid Sunglasses:

  Reflected glare from any smooth surface and scattered glare at midday are both likely to be at least partially polarized *parallel to the ground*. Both are thus blocked by a pair of polaroid sunglasses with a **vertical transmission axis**.

- Doppler Shift, Moving Source:

  In a non-relativistic setting ($v_s \ll c$):

  $$f' = \frac{f}{\left(1 \mp \frac{v_s}{c}\right)} \tag{10.11}$$

  for an approaching (-) or receding (+) source describes the general moving source doppler shift in the frequency/color detected by the receiver.

- Doppler Shift, Moving Receiver:

  Again in a non-relativistic setting ($v_r \ll c$):

  $$f' = f(1 \pm \frac{v_r}{c}) \tag{10.12}$$

  for a receiver moving towards (+) or away from (-) the source.

- Moving Source and Moving Receiver:

  Ditto:

  $$f' = f\frac{(1 \pm \frac{v_r}{c})}{(1 \mp \frac{v_s}{c})} \tag{10.13}$$

- Cerenkov Radiation:

  The "light boom" given off by a charged particle moving faster than the speed of light *in a medium* is called *Cerenkov radiation*.

10.1 The Speed of Light

We just learned that the speed of light in a vacuum, derived from Maxwell's Equations, is $c = 1/\sqrt{\epsilon_0\mu_0} = 3 \times 10^8$ meters/second. However, we have *also* learned that the permittivity and permeability of bulk polarizable matter are not equal to their vacuum equivalents. The conclusion is inescapable. The speed of light is not c in a medium.

We *expect* it to be $v = 1/\sqrt{\epsilon\mu}$ where e.g. $\epsilon = \kappa\epsilon_0$ (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:

$$v \approx \frac{1}{\sqrt{\kappa\epsilon_0\mu_0}} = \frac{c}{\sqrt{\kappa}} \tag{10.14}$$

To keep life simple, we take all of the contributing properties of the material and roll them into a single relation:

$$v_{\text{medium}} = \frac{c}{n} \tag{10.15}$$

n is called the *index of refraction* of the medium and is roughly equal to $\sqrt{\kappa}$.

However, there is a problem with this. κ 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.

10.2 The Law of Reflection

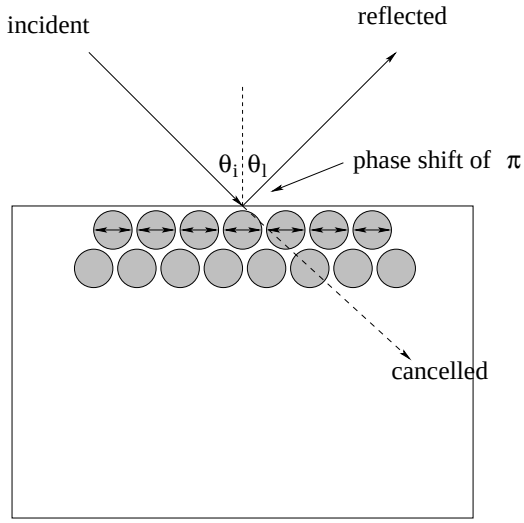


Figure 10.1: When light is incident on a perfectly reflecting surface, it creates little antennas/sources that radiate the *opposite* field in the direction of the incident field. These antennas cause the light to be reflected at the same angle and with the opposite phase from the surface.

A perfect conductor in electrostatic equilibrium, we recall, cancels the electric field inside by arranging charges on its surface to effect the cancellation. Similarly, it creates surface currents that oppose and cancel magnetic fields. In the dynamical case this is still true for good conductors and optical frequencies. An incoming *light* wave strikes the conductor, and its electric field *polarizes* the surface atoms so that they become little antennae that oscillate along with the electric and magnetic field of the light. However, the fields produced *flip over* (the way a dipole field does) and hence propagate in the leading direction with the *opposite phase*, cancelling the forward directed field quite rapidly at the surface (often within a few layers of atoms).

Since the conductor is good, very little energy is lost to eddy current heating during this cancellation. The oscillating surface currents must reradiate their energy, and the only direction they can do so that conserves energy and momentum is to *reflect* the incident energy. However, the reflected wave (in order to achieve the cancellation at the surface) must have the *opposite phase* from the incoming wave. The situation is very much like the reflection of a wave pulse on a string from a fixed point on the wall – the reflected wave flips so it is upside down for precisely the same reasons (energy and momentum conservation).

In an elastic collision with the conductor, the component of the momentum of the light *along* the surface is unchanged, but the perpedicular component inverts (becomes minus itself). The only way this can be true is for the light to bounce off of the surface, with its phase inverted, at an angle of reflection θ_r (measured relative to the normal at the surface at that point) equal to the angle of incidence θ_i as drawn above.

So that's it:

$$\theta_i = \theta_\ell \tag{10.16}$$

is the Law of Reflection. The polarization properties of the reflected light will be discussed later below.

Note well that for this to be strictly true requires that the surface in question be extremely smooth – "shiny" as it were. Otherwise neighboring rays would be reflected at different angles because of small differences in the direction of a normal at different point on a rough surface. Many (even most) surfaces of real materials are indeed rough on a microscopic scale (compared to the wavelengths of the incoming light) and hence are diffusely

illuminated ty light instead of perfectly reflecting it according to this rule. Many materials also differentially absorb light and only "reflect" particular wavelengths and hence colors.

We will assume that the law of reflection holds, more or less perfectly, for shiny smooth good conducting (e.g. metal) surfaces, such as a polished piece of silver or aluminum. This in turn will help us understand how *mirrors* work to form images of objects next week.

10.3 Snell's Law

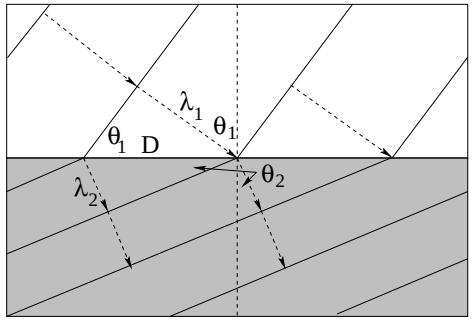


Figure 10.2: When light is incident on a transparent dielectric surface, it is partially transmitted and partially reflected. Since its *speed* changes, however, the light must *change direction* at the surface as shown.

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).

It should be fairly obvious that the *frequency* of light in the two media cannot change. If the same number of wavefronts per second do not pass each point in either medium, wavefronts must be building up in between. This in turn means that energy (associated with the wavefronts) must be building up. This simply does not happen.

It should also be less obvious that the wavefronts themselves – the places where the waves reach their maximum amplitudes – should be the same just inside and just outside the media interface. For it to be otherwise would require a very strange charge distribution on the surface itself, one that one cannot easily imagine arising.

Since the wave must *change speed* across the media interface, and since the speed of the wave is given by:

$$v = \frac{c}{n} = f\lambda \tag{10.17}$$

with the same frequency on both sides, it is clear that the wavelength

$$\lambda = \frac{c}{nf} \tag{10.18}$$

must *also* change, being longer where the speed of light is greater (and n is smaller).

Simple geometry based on these simple ideas requires that the wave will also change direction. We can compute this change and direction from the figure above. If we look at the top triangle with angle θ_1 and hypotenuse D and the bottom triangle with angle θ_2 and the *same hypotenuse* (the distance between wavefronts on the interface between media), we note that:

$$D = \frac{\lambda_1}{\sin(\theta_1)} = \frac{\lambda_2}{\sin(\theta_2)} \tag{10.19}$$

or (substituting from above and cancelling c/f):

$$\frac{1}{n_1 \sin(\theta_1)} = \frac{1}{n_2 \sin(\theta_2)} \tag{10.20}$$

Inverting, we obtain **Snell's Law**:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.21}$$

Since the geometry is exactly the same going from n_2 to n_1, we conclude that it doesn't matter which medium has the greater or the lesser index of refraction.

10.3.1 Fermat's Principle

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.

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.

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.

In figure 10.3, 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

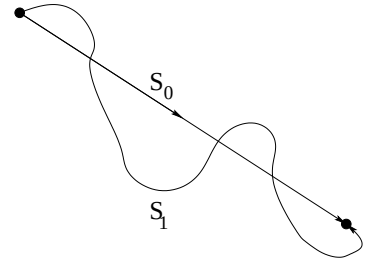


Figure 10.3: For constant speed, the straight line path between A and B takes the least time.

describes this case.

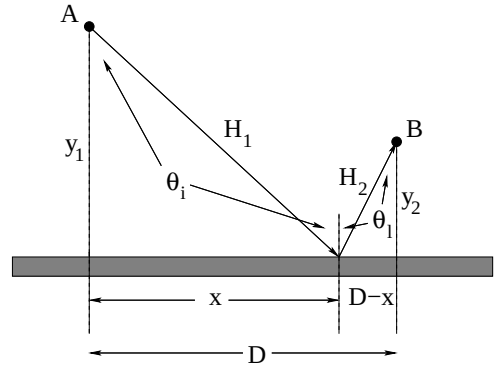


Figure 10.4: The path with $\theta_i = \theta_l$ is the one with the minimal time when the entire trajectory is otherwise in a single medium with a constant speed.

In the figure above we consider reflection. 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:

$$\begin{aligned} H &= H_1 + H_2 \\ &= \sqrt{y_1^2 + x^2} + \sqrt{y_2^2 + (D-x)^2} \end{aligned} \tag{10.22}$$

We need to find a condition that produces the minimum of this function. We therefore differentiate with respect to x, set the result to zero, and solve for (say) x or θ_1. y_1, y_2 and D are all constant, so:

$$\frac{dH}{dx} = \frac{2x}{\sqrt{y_1^2 + x^2}} - \frac{2(D-x)}{\sqrt{y_2^2 + (D-x)^2}} = 0 \tag{10.23}$$

or

$$\sin(\theta_i) = \frac{2x}{\sqrt{y_1^2 + x^2}} = \frac{2(D-x)}{\sqrt{y_2^2 + (D-x)^2}} = \sin(\theta_l) \tag{10.24}$$

If the speed of light is a constant, this condition minimizes both distance and hence time $t = H/v$. Thus $\theta_i = \theta_l$, and we see that the Law of Reflection is consistent with Fermat's principle as well.

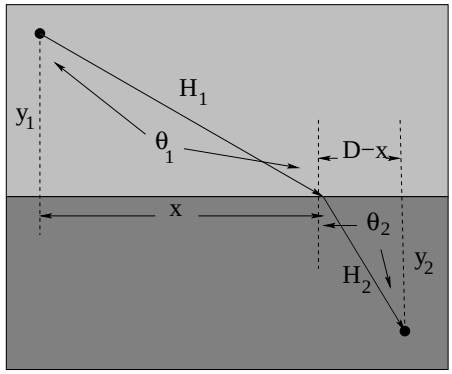


Figure 10.5: The path with $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$ is the one with the minimal time when the trajectory goes between media n_1 and n_2 where light has distinct speeds. As suggested, one minimizes the time by choosing a trajectory that trades off more distance in the faster medium against less distance in the slower one.

As before, we only need consider straight line trajectories *in* a given medium, and so the figure above is all we need to consider.

This time, since the speeds in the two media are *different*, we have to directly optimize the time. We form:

$$t_1 = \frac{\sqrt{y_1^2 + x^2}}{v_1} = \frac{n_1\sqrt{y_1^2 + x^2}}{c} \tag{10.25}$$

and

$$t_2 = \frac{\sqrt{y_1^2 + (D-x)^2}}{v_2} = \frac{n_2\sqrt{y_1^2 + (D-x)^2}}{c} \tag{10.26}$$

as the *time* taken for the light to travel in a straight line from A to x and from x to B.

The total time is thus:

$$t = t_1 + t_2 == \frac{n_1\sqrt{y_1^2 + x^2}}{c} + \frac{n_2\sqrt{y_2^2 + (D-x)^2}}{c} \tag{10.27}$$

Differentiating and setting the result equal to zero is the *same* algebra as above, except that there is an extra factor of n_1 and n_2 on each side. The details are left as a (simple) exercise; the result is:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.28}$$

and we see that Snell's law is consistent with 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 the action!

10.3.2 Total Internal Reflection, Critical Angle

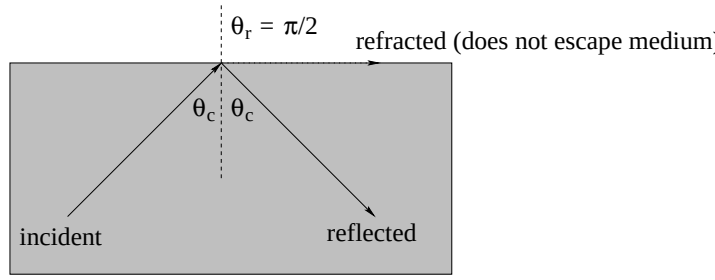


Figure 10.6: Light travelling from a denser medium to a lighter one is totally internally reflected if $\theta_i \geq \theta_c = \sin(\frac{n_1}{n_2})$, corresponding to an angle of refraction of $\pi/2$, where the refracted ray *fails to escape the medium*.

If a ray is travelling from a denser medium to a lighter one, one quickly observes a curious thing. Since the ray is bent *away* from the normal, there exist angles for which Snell's law has no solution!

In fact, it is easy to identify an angle of incidence such that the angle of refraction is $\theta_r = \pi/2$. If we assume that $n_2 > n_1$ and we are going from medium n_2 (the heavier/denser) to medium n_1 (the lighter/less dense):

$$n_2 \sin(\theta_2) = n_2 \sin(\theta_c) = n_1 \sin(\pi/2) = n_1 \tag{10.29}$$

or

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right) \tag{10.30}$$

If we increase $\theta_2 > \theta_c$, we make the left hand side of Snell's law *bigger than* n_1 but we cannot find any angle θ_r for which $\sin(\theta_r) > 1$!. We conclude that at all angles θ_c and greater the ray *fails to escape the medium!*

Since it is not absorbed by the interface, and is not transmitted into medium n_1, the only place the energy in this ray can go is into the *reflected* ray. The ray is thus *totally internally reflected*.

Total internal reflection is extremely useful in our modern society. It is the basis of *fiber optics* where (laser) light signals are "trapped" inside a "light pipe" that transmits the light down the fiber and around sufficiently gentle bends without allowing the light to escape through the sides of the optical fibers that have an index of refraction greater than that of the surrounding air or other media.

It is also pretty! Diamonds and the diamond-like compound C3 (Moissonite) have *extremely large indices of refraction*, roughly $n_d = 2.4$. This makes its critical angle:

$$\theta_{cd} = \sin^{-1}\left(\frac{1}{2.4}\right) = 24.6^\circ \tag{10.31}$$

Light incident on the facet of a diamond at any angle *greater* than this (rather small) angle is *trapped* by the diamond. Diamonds are cut so that light entering through any given facet is reflected many times without escaping, so that dispersion splits the light up into many colors until it escapes either through the sides or at corners or edges. This gives diamond (or Moissanite) its "bright and sparkly" appearance. Cut crystal prisms and lesser clear gemstones have much the same properties on a lesser scale, trapping light and splitting it up into a rainbow of colors to brighten an otherwise drab existence.

10.4 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 polaraized

light.

10.4.1 Unpolarized Light

Unpolarized light is light for which the polarization vector is constantly shifting its direction around. For a few tens to thousands of wavelengths the electric field vector points in some direction. Then it suddenly shifts into a new direction, as its source gets randomly interrupted. Unpolarized light is typically produced by "hot" or "random" sources such as the Sun, a hot lightbulb filament, the gas in a fluorescent bulb, a candle flame. On average, unpolarized light has its energy/intensity equally distributed between the two independent directions of polarization.

10.4.2 Linear Polarization

Linear polarization occurs whenever the electric field vector oscillates consistently in a single vector direction in the plane perpendicular to propagation. The following are all examples of linearly polarized light propagating in the z-direction with frequency ω:

Light linearly polarized in the x-direction:

$$\boldsymbol{E}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega t) \tag{10.32}$$

(The associated magnetic field *must* be:

$$\boldsymbol{B}(z,t) = B_{0y}\hat{\boldsymbol{y}}\sin(kz - \omega t) = \frac{E_{0x}}{c}\hat{\boldsymbol{y}}\sin(kz - \omega t) \tag{10.33}$$

according to the rules derived in the previous chapter, because

$$|\boldsymbol{B}| = \frac{|\boldsymbol{E}|}{c} \tag{10.34}$$

and because

$$\hat{\boldsymbol{x}} \times \hat{\boldsymbol{y}} = \hat{\boldsymbol{z}} \tag{10.35}$$

in the Poynting vector.)

Light linearly polarized in the y-direction:

$$\boldsymbol{E}(z,t) = E_{0y}\hat{\boldsymbol{y}}\sin(kz - \omega t) \tag{10.36}$$

(The associated magnetic field *must* be:

$$\boldsymbol{B}(z,t) = -B_{0x}\hat{\boldsymbol{x}}\sin(kz-\omega t) = -\frac{E_{0y}}{c}\hat{\boldsymbol{x}}\sin(kz-\omega t) \tag{10.37}$$

according to the rules derived in the previous chapter, because

$$\hat{\boldsymbol{y}} \times -\hat{\boldsymbol{x}} = \hat{\boldsymbol{z}} \tag{10.38}$$

in the Poynting vector.)

Finally, light linearly polarized along the line at $\pi/4$ above the x-axis is::

$$\boldsymbol{E}(z,t) = \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz-\omega t) + \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz-\omega t) \tag{10.39}$$

The amplitude of the electric field is E_0 (why?). What must the direction and magnitude of the associated magnetic field?

10.4.3 Circularly Polarized Light

There is no reason that the magnitudes of the electric polarization components in the two independent directions have to be *the same* or to be *in phase*. We start by considering the case where they have the same magnitude but are $\pi/2$ out of phase:

$$\begin{aligned}\boldsymbol{E}(z,t) &= \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz-\omega t \pm \pi/2) + \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz-\omega t) \\ \boldsymbol{E}(z,t) &= \frac{\sqrt{2}}{2}E_0\left(\pm\hat{\boldsymbol{x}}\cos(kz-\omega t) + \hat{\boldsymbol{y}}\sin(kz-\omega t)\right)\end{aligned} \tag{10.40}$$

These two components describe a vector of constant length that sweeps around in a *circle*, either counterclockwise (-) or clockwise (+). We call this *circularly polarized light*. Note that the two components must have equal amplitudes and must be $\pi/2$ out of phase to be circularly polarized. There are two independent *helicities* of circularly polarized light: right (clockwise/+) and left (anticlockwise/-) when facing *in* the direction of propagation).

10.4.4 Elliptically Polarized Light

If the amplitudes of the two waves are (potentially) different *and* the two waves are (potentially) out of phase, the most general polarization state is

that of *elliptical* polarization:

$$\boldsymbol{E}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega t + \delta_x) + E_{0y}\hat{\boldsymbol{y}}\sin(kz - \omega t + \delta_y) \tag{10.41}$$

In this expression, E_{0x} and E_{0y} may or may not be equal, and the phases δ_x and δ_y may or may not be zero *or* equal. The amplitudes of the x and y limits define a rectangular box. The electric field vector rotates within that box wit the box tipped at an angle relative determined by the relative phase difference $\delta = \delta_x - \delta_y$ (where if $\delta = 0$ or $\delta = \pi$ one has linear polarization).

To see a lovely animation of the electric field vector for various flavors of polarization, visit:

http://www.nsm.buffalo.edu/∼jochena/research/opticalactivity.html

10.4.5 Polarization by Absorption (Malus's Law)

A polaroid filter is made by putting oriented conducting threads into a transparent medium in such a way that long currents in those threads created by the polarization component of light parallel to the thread heats the threads, absorbing and attenuating *only* that component of the incident polarized or unpolarized light and passing the component perpendicular to the threads (the **transmission axis** of the filter).

The rules for transmission are simple. If the incident light is unpolarized, on average half its energy is polarized in either polarization direction. Therefore (assuming that the filter is "ideal" and otherwise fully transparent):

$$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \tag{10.42}$$

The transmitted light is fully linearly polarized in the direction of the transmission axis of the filter.

If the light that is incident on the filter is already polarized, then only the *component* of the electric field vector that is *parallel* to the transmission axis is transmitted. That is:

$$E_{\text{transmitted}} = \boldsymbol{E}\cdot\hat{\boldsymbol{t}} = E_{\text{incident}}\cos(\theta) \tag{10.43}$$

where θ is the angle between the direction of linear polarization of the incident light and a unit vector along the transmission axis.

To find the transmitted *intensity*, we need just remember the relation between the electric field strength and the intensity that follows from the intensity being the time-average magnitude of the Poynting vector:

$$I = \left| \frac{1}{2\mu_0} \boldsymbol{E} \times \boldsymbol{B} \right| = \frac{1}{2\mu_0 c} E^2 \tag{10.44}$$

The intensity is directly proportional to the electric field amplitude, squared, so that:

$$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \tag{10.45}$$

This result is known as **Malus's law**.

10.4.6 Polarization by Scattering

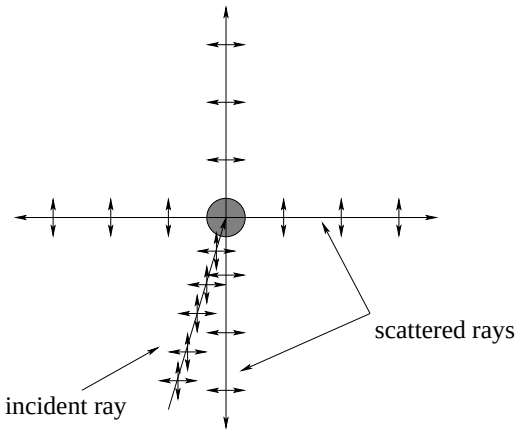


Figure 10.7: The scattering of initially unpolarized light by a molecule or dust particle. Note that the polarization is perpendicular to the *plane of scattering* for each of the possible outgoing directions.

When unpolarized light passes across an atom or molecule, it *polarizes* it in the instantaneous direction of the electric field vector (which, recall, has a definite direction at any time but which jumps around to a new direction every 10-1000 optical periods). The oscillating molecule acts like a *dipole antenna* and *reradiates* the incident electromagnetic wave. However, the reradiated electric field must be *parallel* to the dipole moment of the molecule, and there is no radiation *along* the dipole (with a clear maximum at right angles to the dipole. As a consequence we can easily see that the

rule for polarization of rays scattered more or less at right angles is that they must be polarized *perpendicular to the plane of scattering!*

10.4.7 Polarization by Reflection

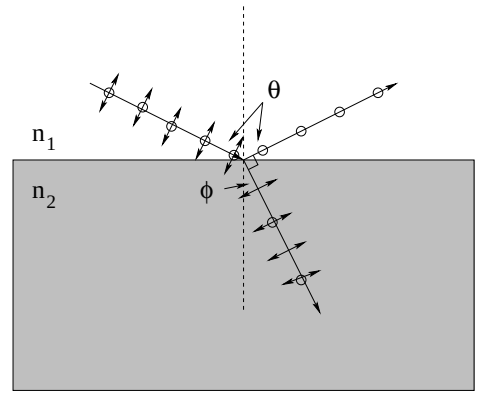


Figure 10.8: The scattering of initially unpolarized light by reflection off of a plane surface between two dielectric media at the *Brewster angle* that produces complete polarization of the reflected ray. Note that the polarization of all reflected rays incident on the surface at an angle is *parallel to the ground* even at angles other than the Brewster angle.

When light strikes a surface between two regions with differing indices of refraction, it is partially transmitted and partially reflected (with the amount of each determined by the angle of incidence and the two indices of refraction). The reflection is caused by the polarization of surface molecules in such a way that the light scattered by them adds up coherently into the reflected wave; similarly those polarized molecules create a forward propagating wave into the medium (although at a different angle according to Snell's law). As before, the polarized surface molecules (dipoles) *cannot radiate along their own axis* so that light that is reflected *parallel* to one of the polarization directions cannot contain that polarization.

This state of affairs occurs when the reflected ray is perpendicular to the refracted ray, pictured above. In this case:

$$n_1 \sin(\theta) = n_2 \sin(\phi) \tag{10.46}$$

is Snell's law, but clearly:

$$\phi = \frac{\pi}{2} - \theta \tag{10.47}$$

so that:

$$\sin(\phi) = \sin(\pi/2 - \theta) = \cos(\theta) \tag{10.48}$$

and *Brewster's formula*:

$$\tan(\theta_b) = \frac{n_2}{n_1} \tag{10.49}$$

is the condition for θ_b, the so-called *Brewster angle* of incidence (and hence reflection) where the reflected ray is completely polarized parallel to the surface (and perpendicular to the plane of reflection, just as was the case with scattered light above).

However, the polarization component in the plane of reflection is *always* reduced at angles other than $\theta = 0$ as the component of the polarization gradually lines up with the reflected ray so reflected light is at least *partially* polarized in the plane at all angles other than 0. Note that the transmitted light is *partially* polarized *in* the plane of transmission – this is not complete because all of the perpendicularly polarized light is not reflected at the surface, some is still transmitted into the medium.

10.4.8 Polaroid Sunglasses

As we have just seen, reflected glare from any smooth surface is likely to be at least partially polarized parallel to the ground. It is thus blocked by a pair of polaroid sunglasses with a *vertical* transmission axis. Similarly, (scattered) light from the blue sky viewed near the horizon at midday is predominantly polarized parallel to the ground and is *also* blocked by a vertical transmission axis, which can make e.g. driving safer and less stressful on the eye.

10.5 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.

The idea underlying the Doppler shift is very simple. If the source is moving towards the receiver, its motion foreshortens the normal wavelength, increasing the frequency observed by the stationary receiver. If the receiver is moving towards the source, its motion reduces the time between the wavefronts it receives, increasing the frequency it observes. If both motions are occurring, both shifts occur as a product. We show the picture and quick derivation of each possibility below.

10.5.1 Moving Source

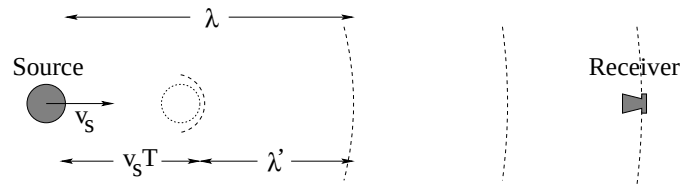


Figure 10.9: Wave geometry for Doppler shift of moving source.

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:

$$\begin{aligned} \lambda' &= \lambda - v_s T \\ cT' &= cT - v_s T \\ T' &= T\left(1 - \frac{v_s}{c}\right) \\ \frac{1}{T'} &= \frac{1}{T}\frac{1}{\left(1 - \frac{v_s}{c}\right)} \\ f' &= \frac{f}{\left(1 - \frac{v_s}{c}\right)} \end{aligned} \tag{10.50}$$

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:

$$f' = \frac{f}{\left(1 \mp \frac{v_s}{c}\right)} \tag{10.51}$$

for an approaching (-) or receding (+) source describes the general moving source doppler shift in the frequency/color detected by the receiver.

Note well that visible light sources moving away from the receiver are shifted towards the *red* end of the spectrum, while sources moving towards the receiver are shifted towards the *violet* end of the spectrum. Since spectral lines produced by atoms have sharp and well-defined frequencies, this permits us to ascertain that the visible Universe is *expanding* (as all distant stars and galaxies are red-shifted). Since the velocity with which distant stars are receding from the Earth *increases with distance*, the red shift becomes a *meter stick* permitting us to measure the size of the visible Cosmos. This is a small but significant part of the physical evidence for the *Big Bang* cosmological model that so far seems best to fit the data, and that suggests that the Big Bang occurred approximately 13.5 billion years ago (give or take a billion years) so that the visible Cosmos is a sphere roughly 27 billion light years across, containing roughly a trillion galaxies containing order of a trillion stars apiece. This is around Avogadro's number of *stars*.

With no boundaries visible in any direction, there is no particular reason for us to think that we are in the exact center of the cosmos, save in the sense that every point is in the middle of an infinite line. Sometimes small pieces of physics (such as the Doppler shift of light) can have *enormous* consequences.

10.5.2 Moving Receiver

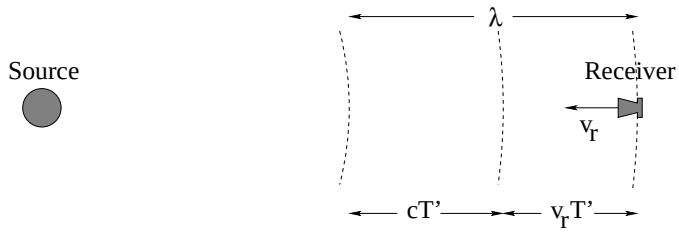


Figure 10.10: Wave geometry for Doppler shift of a moving receiver.

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 $\lambda = cT = (c + v_r)T'$:

$$\begin{aligned} cT &= (c + v_r)T' \\ T &= (1 + \frac{v_r}{c})T' \\ \frac{1}{T'} &= \frac{1}{T}(1 + \frac{v_r}{c}) \\ f' &= f(1 + \frac{v_r}{c}) \end{aligned} \tag{10.52}$$

As before, if the receiver is moving away, it decreases f' instead of increasing it, so that the general rule is:

$$f' = f(1 \pm \frac{v_r}{c}) \tag{10.53}$$

for a receiver moving towards (+) or away from (-) the source.

10.5.3 Moving Source and Moving Receiver

The rule is just the product of the two rules:

$$f' = f\frac{(1 \pm \frac{v_r}{c})}{(1 \mp \frac{v_s}{c})} \tag{10.54}$$

It is interesting to note that if a source is moving at the speed of light (where these expressions are no longer valid, alas, 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.

10.6 Homework for week 10

Problem 1.

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 2.

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 occurence of tornados, and is a technology used in a variety of medical imaging techniques including e.g. ultrasound.

Problem 3.

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 4.

Derive Brewster's Formula (the expression for the angle of incidence for which reflected light is completely polarized parallel to the surface).

Problem 5.

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 6.

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 7.

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 8.

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.)

Problem 9.

Until I have time to absorb them in this book, do problems 78 and 84 out of Tipler and Mosca (chapter 31).

Week 11: Lenses and Mirrors

- The distance from a mirror (or lens) to an object one is viewing in (or through) it is s, the **object distance**. Object distances are positive if the object is on the side of the mirror (or lens) that the light is coming *from*. Object distances are obviously 'always' positive, unless the object is a *virtual object* formed out of the image of a previous mirror or lens, which can be either positive or negative.

- The distance from a lens or mirror to the image one is viewing is s', the **image distance**. Image distances are positive if the image is on the side of the mirror (or lens) that the light is going *to*.

- The focal length f of a mirror (or lens) is the point where incident parallel rays are focused **to** (for positive focal lengths) or appear to be defocused **from** (for negative focal lengths). f is typically measured in meters (SI) or centimeters (for convenience). However, the strength of *lenses* is usually given in *diopters*, where:

$$d = \frac{1}{f} \tag{11.1}$$

  with f in meters. This a one diopter (1.00d) lens has a focal length of 1 meter. A 10.00d lens has a focal length of 0.1 meter. A diverging lens with a focal length of one centimeter is -100.00d.

- The mirror (or thin lens) equation relating s, s', and f is:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{11.2}$$

- The transverse magnification of a simple mirror (or lens) is defined by the ratio of the image height y' to the object height y:

$$m = \frac{y'}{y} = -\frac{s'}{s} \tag{11.3}$$

- A **real image** is one where the rays of light that appear to the eye to diverge from a point on the image actually pass through that point. A **virtual image** is one where the rays of light that appear to the eye to diverge from a point on the image do *not* actually pass through the image.

- In addition to being real or virtual, an image can be **erect** (oriented the same way as the object) or **inverted** (oriented the opposite way from the object.

- For a spherical mirror, the focal length is given by:

$$f = \frac{r}{2} \tag{11.4}$$

  where r is positive when it is on the side of the mirror reflected light is going *to*.

- For a thin lens, the focal length is given by the **lensmaker's formula**:

$$\frac{1}{f} = (n_2 - n_1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{11.5}$$

  In this expression, n_1 is the index of the surrounding medium (typically air, $n_1 = 1$) and n_2 is the index of refraction of the lens itself. r_1 (r_2) is the radius of curvature of the *first (second) surface struck* by the ray, with the sign convention that it is positive (negative) on the side of the lens refracted light is going *to* (coming *from*).

  The advantage of using diopters as a measure of lens strength is inherent in this expression, as you can see that the combined strength of the two lensing surfaces (in diopters) is equal to the *sum* of the strength of *each* surface, in diopters. This extends to any pair of lenses placed close together – the effective strength of two lenses closely placed (relative to their focal lengths) in front of one another is the sum of their strength in diopters.

- True Facts about the Eye:

  The eye is approximately one inch in diameter. A *lens* in front casts a *real* image of objects being viewed onto its *retina*, where rods and cones transform the light into neural impulses which are then conveyed to the brain for processing by the optic nerve. Rods and cones are very

sensitive to light (and easily damaged) – the light content is regulated by the *iris* of the eye, which expands and contracts the *pupil* – the aperture through which light passes as it enters the lens.

The focal length of a relaxed lens of an eye with *normal* vision is on the retina, so distant objects are automatically in focus. Given the diameter of the eye, this means that the strength of the lens of a normal eye is approximately 40.00d. The focal length of a relaxed *farsighted* eye is *behind* the retina (too long, strength less than 40.00d) and is corrected with a *converging* lens to make up the difference. The focal length of a relaxed *nearsighted* eye is in *front* of the retina (too short, strength greater than 40.00d) and is corrected with a *diverging* lens to take away some of its strength.

There are muscles that surround the lens of the eye in a ring that contract, making the lens bulge (to a greater radius of curvature) and thereby *shortening* the focal length (a process called *accommodation*) to bring nearby objects into focus. The nearest point one can bring an object to the eye and still bring it into focus on the retina is called the *near point* of the eye and is also the *distance of most distinct vision*, represented x_{np}. In most adults, this distance is around 25 cm (less for small children, longer for the elderly).

A nearsighted person's lens *already* has too short a focal length to be able to focus distant objects on the retina, and accommodation only shortens the focal length still farther. A nearsighted person cannot see anything clearly at distances *greater* than some point, called the *far point* for that person's eyes. A nearsighted person is one for whom the far point x_{fp} is less than infinity.

- The simple magnifier is a converging ($f > 0$) lens placed immediately in front of the eye. An object placed at its focal point therefore forms a virtual image at infinity that is automatically brought into focus by the relaxed normal (or vision corrected) eye. The magnification of the object occurs because one can bring the object *closer* to the eye than x_{np} and still see it clearly, where it subtends a *greater* angle on the retina (angular magnification). Its magnification is given by:

$$M = \frac{x_{np}}{f} \tag{11.6}$$

It is very important to understand the simple magnifier, as it forms the eyepiece of *both* the microscope *and* the telescope.

- A telescope is used to view a distant object by making the angle its image subtends on the retina larger. Two lenses are situated at ends of a tube such that their focal points are coincident. The first lens (with a long focal length) forms a *real image* of the distant object more or less at its focal point. The second lens (with a short focal length) is used to view this real image as a simple magnifier. This produces a virtual image at infinity that subtends a greater angle than the original object did, viewable with the relaxed normal eye.

  The overall angular magnification of a telescope is given by:

$$M = -\frac{f_o}{f_e} \tag{11.7}$$

  The eyepiece lens can be converging (regular) or diverging (Galilean). In both cases this formula for the magnification works (provided that one uses a negative f_e for the diverging lens and place the focal point f_o at the focal point on the *far* side of the diverging lens). A regular telescope inverts the image, which is inconvenient and undesireable. A Galilean telescope does not invert the image.

- A compound microscope is used to view a very small, but nearby object. It accomplishes this in two stages. Two short focal length lenses are situated at ends of a tube much longer tube. The *tube length* ℓ of the microscope is by definition the distance between the focal point of the first, or *objective* lens (which must be converging) and the second, or *eyepiece* lens. The object is placed just outside of the focal length of the objective lens in such a way that it forms a *magnified, real image* of the object more or less at the end of the tube length. The eyepiece lens is used as a simple magnifier to view this real image, and can be converging or diverging as was the case for the telescope. It produces a virtual image at infinity that subtends a greater angle than the real image formed by the objective lens alone would if viewed at the near point of the relaxed normal eye.

  The magnification of the objective is:

$$M_o = -\frac{\ell}{f_o} \tag{11.8}$$

The magnification of the eyepiece (simple magnifier) is:

$$M_e = \frac{x_{np}}{f_e} \tag{11.9}$$

The overall magnification is therefore:

$$M_{tot} = -\frac{\ell\ x_{np}}{f_o f_e} \tag{11.10}$$

where as before, this formula for the magnification works provided that one uses a negative f_e for the diverging lens and place the real image formed by the objective on the *far* side of the diverging lens. A regular microscope inverts the image, which is inconvenient and undesireable. A "Galilean" microscope does not invert the image.

11.1 Vision and Plane Mirrors

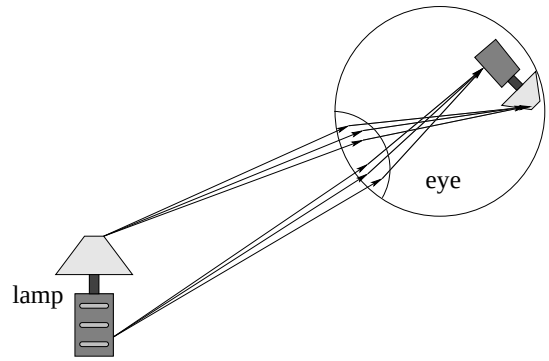


Figure 11.1: How the eye sees an object. Light diverging from points on the surface of the object are focused onto the retina of the eye, where they form an *image* of the object that the retina converts into neural impulses and your brain converts into perception.

Objects in the real world that are illuminated by diffuse light absorb the light at every point on their surface and then reradiate (selected colors/frequencies) from each point in all directions. This is why you can see something that is illuminated from all angles – every point on its surface emits light reradiated from the illuminating source in all directions so no matter where you look at it from, some of the light reaches your eye.

To *completely* understand how your eye can see the object, we have to get halfway through this week's work. On the other hand, we can't understand enough about how mirrors and lenses work to understand the eye without understanding the eye well enough to understand how lenses and mirrors work.

Hmmm, a bit of a dilemma. We have to *bootstrap* just a bit and draw a few pictures now that you won't completely understand later to help you understand what you need to understand what you need to understand later. Or something like that.

So meditate on the picture above, which shows light diffusely scattered from from a couple of points on a common object. The light goes in *all directions* from *all of the points on the surface of the object.* Some of these rays reach your eye. There the lens of your eye does its thing, and forms a nice sharp *image* of the object cast upon the retina of the eye. Vision occurs.

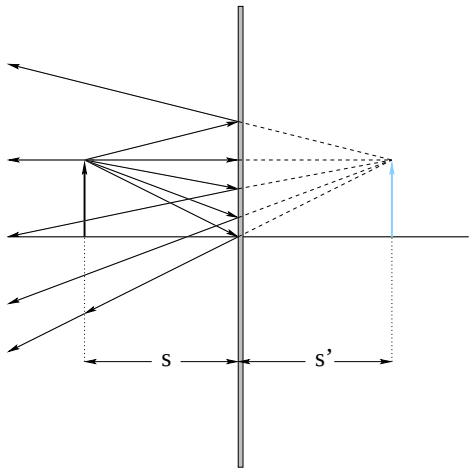


Figure 11.2: The geometry of forming an image in a plane mirror.

Now consider looking at an object in a *plane mirror.* Lamps are too hard to draw, so we consider an arrow, which we will use as a "generic object" in our diagrams.

Rays radiated from the object radiate out in all directions as shown in the figure above. When they strike the mirror they are reflected with the angle of incidence equal to the angle of reflection. As we look at the mirror, we see the rays that originated on a single point on the object *as if* they

were diverging from a single point in space. That point is the *image* of the point on the object. Since every (visible) point on the object corresponds to an apparent point of divergence in space from the image, we can see the image *exactly as if* we were looking at an object.

In the case of a plane mirror (above) the image is always *behind* the mirror. The light rays you see do not actually pass through the image, they simply appear to diverge from it. We call such an image a *virtual* image.

We need to define several quantities that will be essential in our analysis of how lenses and mirrors work. The distance from a mirror (or lens) to an object one is viewing in (or through) it is s, the *object distance*. Object distances are *positive* if the object is on the side of the mirror (or lens) that the light is coming *from*. Object distances are obviously 'always' positive, unless the object is a *virtual object* formed out of the image of a previous mirror or lens, which can be either positive or negative.

The distance from a lens or mirror to the image one is viewing is s', the *image distance*. Image distances are *positive* if the image is on the side of the mirror (or lens) that the light is going *to*.

Multiple mirrors can be used to create images of images, or images of images of images (used as "virtual objects" for the second mirror). Most of us have experienced the "infinite tunnel" of images that results from standing directly in between two plane mirrors.

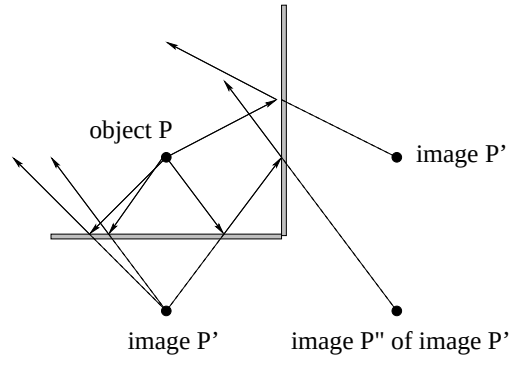


Figure 11.3: Two mirrors create an image of an image. Only a few of the many rays are drawn – copy the picture and fill in more yourself.

11.2 Curved Mirrors

Plane mirrors simply create a perfect image of everything that is in the real space reflected in the mirror. Things get more interesting if the mirrors are *curved*. Curved mirrors can create images that are systematically larger or smaller than the object, and can create a new kind of image from the one seen in figure (11.2).

In figure (11.4) we see a *concave spherical mirror*, which we will also call a *converging* mirror or a *positive* mirror[1]. The horizontal line running through the center of the mirror is very important and is called the *axis* of the mirror, which is rotationally symmetric about this axis. Even imaging an arrow is too complicated for our purpose (which is to figure out how spherical mirrors can make images at all) so we look for the image of a *single point* P, which we locate for convenience on the axis of the mirror.

The image P' occurs where two reflected rays *cross*. The two rays in question are the one that strikes a distance l up the mirror (with angle of incidence equal to the angle of reflection) and a ray that goes along the axis and is reflected directly back the way it came. This is a new kind of image – the rays don't just *appear* to come from a point in space (a point that is really in the dark of your closet or medicine cabinet, back behind the mirror) as they do with a virtual image, they *really* reach the eye after passing through a point in space. You could reach out and put your finger through the point in space they appear to be coming from. We call this kind of image a **real image**, and we need to be able to determine whether an image is real (the kind of image that can be projected on a retina, piece of film, wall, projector screen) or virtual (which cannot be projected at all, since no light actually passes through the image), so be sure you understand the distinction and can categorize images you determine from e.g. ray diagrams.

We begin by making an essential approximation. We will later talk about *aberrations* of 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

[1]For those who have concave/convex dyslexia, remember that con*cave* is like a cave, and curves inward, while con*vex* is nothing at all like a vex. What is a vex, anyway?

image. We can't have that, so we insist that the rays we will deal with be *paraxial* – close to the axis and close to parallel. The former means that we strike the mirror close enough to its center for us to be able to pretend that the deflection occurs in a (slightly) curved plane; the latter means that small angle approximations will all work quite well.

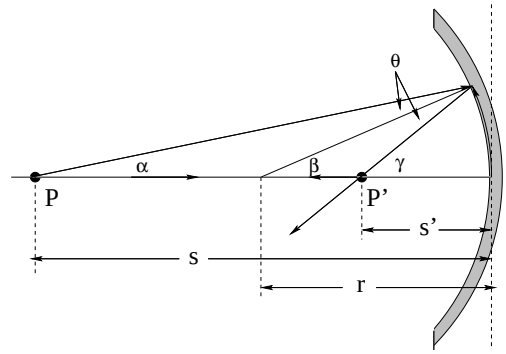


Figure 11.4: The geometry of forming an image in a concave mirror.

Three important lengths are drawn onto the figure: s, s', and r, as well as the distance l itself. Note well also the four angles: α, β, γ and the angle of incidence/reflection θ. Since the angles are all small and l is close to a straight line:

$$\alpha \approx \frac{l}{s} \tag{11.11}$$

$$\beta = \frac{l}{s} \tag{11.12}$$

$$\gamma \approx \frac{l}{s'} \tag{11.13}$$

(where the result for β, note well, is exact because l really is the length of a circular arc that is subtended by the angle β).

We now play games with the triangles in the picture. We use the following rule several times: Consider the triangle with α, θ and the angle δ (filled in to figure (11.5)). We can easily see that $\alpha + \theta + \delta = \pi$. But we can *also* see that $\delta + \beta = \pi$. Therefore:

$$\alpha + \theta = \beta \tag{11.14}$$

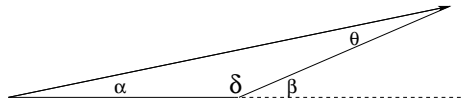


Figure 11.5: $\alpha + \theta = \beta$.

and similarly (considering the other triangle involving β and θ)

$$\beta + \theta = \gamma \tag{11.15}$$

If we eliminate θ, we get:

$$\alpha + \gamma = 2\beta \tag{11.16}$$

Finally, if we substitute in all of the small angle approximations and cancel l, we get:

$$\frac{1}{s} + \frac{1}{s} = \frac{2}{r} \tag{11.17}$$

As we move the object back farther and farther from the mirror (let $s \to \infty$) we note that the image distance approaches $r/2$. Rays coming from an infinitely distant object arrive at the mirror *parallel* and converge *at* $s' = r/2$. We *define* the point where a lens or mirror focuses *parallel, paraxial rays* to be the **focal point** of the lens or mirror. Thus:

$$f = \frac{r}{2} \tag{11.18}$$

and

$$\frac{1}{s} + \frac{1}{s} = \frac{1}{f} \tag{11.19}$$

This is a *very important result!* It is the equation we will use to analyze *all images formed by curved mirrors and thin lenses* (after we derive the same formula for the latter) so be sure that you have learned it and understand it.

The focal length f of a mirror (or lens) is the point where incident parallel rays are focused **to** (for positive focal lengths) or appear to be defocused **from** (for negative focal lengths). f is typically measured in meters (SI) or centimeters (for convenience). However, the strength of *lenses* is usually given in *diopters*, where:

$$d = \frac{1}{f} \tag{11.20}$$

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:

$$x + x' = d \tag{11.21}$$

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.

11.3 Ray Diagrams for Ideal Mirrors

To construct our ray diagrams, we need to begin by idealizing spherical mirrors in a way that "hides" things like the fact that many rays we might wish to image with are *not* paraxial. Later in this chapter we'll deal with many of the aberrations that are features of real lenses and mirrors as deviations from ideal behavior in the focussing elements themselves or the light that

goes through them, but these will be "corrections" that should not cloud our perception of how things basically work.

First, when drawing rays in a ray diagram, one always assumes that *all deflection by the lens or mirror occurs in a single plane.* This is an idealization, to be sure – the reason mirrors and lenses focus light is because they are *curved*, not planar. But paraxial rays by definition strike close enough to the center that the deviation from planar can be ignored, and we idealize this to the entire plane.

Given this, the following three rays have rules that can be used to locate images and compute magnification for any mirror (and eventually, lens):

1. **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**.

2. **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.*

3. **The Central Ray:** A ray from the object that strikes the mirror in the center is reflected by the mirror **with angle of incidence equal to the angle of reflection** which means that the reflected ray is symmetric across the axis from the incident one.

Now consider the following ray diagrams for various positions of our archetypical arrow object for converging (+) and diverging (-) ideal mirrors.

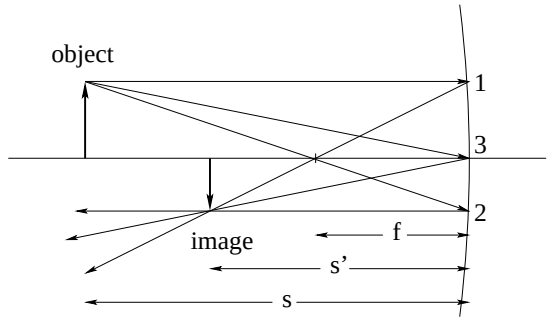


Figure 11.6: Converging mirror with $s = 25 > f = 10$.

In this figure, $f = 10$ cm, $s = 25$ cm. Therefore:

$$\frac{1}{25} + \frac{1}{s'} = \frac{1}{10}$$

$$\begin{aligned} \frac{1}{s'} &= \frac{1}{10} - \frac{1}{25} \\ \frac{1}{s'} &= \frac{1.5}{25} \\ s' &= \frac{25}{1.5} = 16.7 \text{ cm} \end{aligned} \tag{11.22}$$

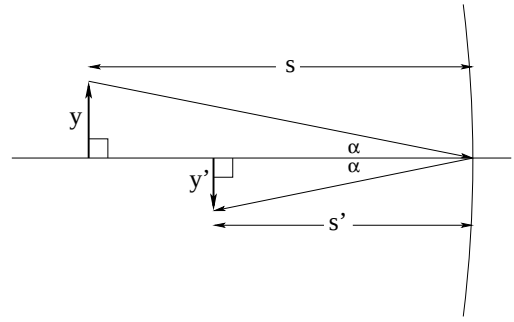


Figure 11.7: Transverse magnification can be determined from the two right triangles formed with the central ray as a hypoteneuse.

To compute the magnification of the image formed above, we note that:

$$\tan(\alpha) = -\frac{y}{s} = \frac{y'}{s} \tag{11.23}$$

(where we rigorously follow the convention that counterclockwise rotation is positive to assign the signs). We define the transverse magnification m of a simple mirror (or lens) is defined by the ratio of the image height y' to the object height y. If we rearrange the terms in this expression, we obtain:

$$m = \frac{y'}{y} = -\frac{s'}{s} \tag{11.24}$$

This expression is valid for all images obtained for any ideal lens or mirror.

Note that in this case, the image formed is real (because the light rays pass through the actual object), inverted, and that the image formed is smaller than the original object.

Let's look at two more possibilities for converging/concave mirrors. In figure (11.8), we see an (upside down) object at a position between f and $2f$. This range is the second possibility for this kind of mirror, one that leads to a *magnified* real image larger than the object.

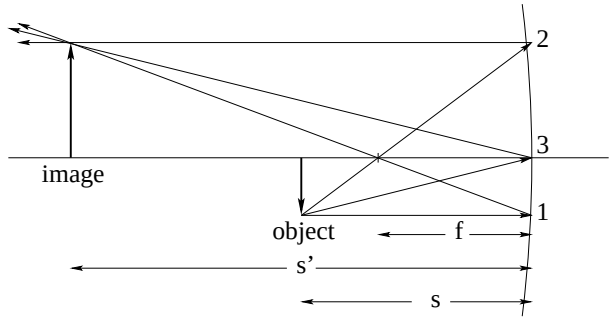


Figure 11.8: Converging mirror with $2f = 20 > s = 15 > f = 10$.

As before, $1/s' = 1/10 - 1/15 = 1/30$ so $s' = 30$ cm. The magnification is $m = -'s/s = -30/10 = -3$. The image is again real and inverted (relative to the object), but in this case the image is larger than the object.

Note that for $s > f$ there is a *symmetry* between solutions with $s > 2f > s'$ and solutions with $s' > 2f > s$, emphasized in the figure above by deliberately drawing the object upside down so that it looks very much like figure (11.8). In fact *any* ray diagram involving real images can work both ways, with s and s' (and the role of the object and image) interchanged because $1/s$ and $1/s'$ appear symmetrically in the mirror/thin lens equation.

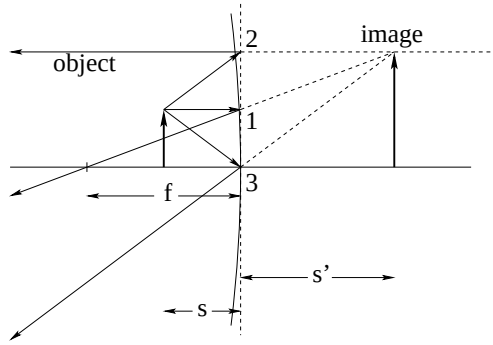


Figure 11.9: Converging mirror with $s = 5 < f = 10$.

In figure (11.9) the third and last distinct possibility for a converging mirror is drawn. In this case, the object is located *inside* the focal length at $s = 5$ cm (for $f = 10$ cm). Thus $1/s' = 1/10 - 1/5 = -1/10$ or $s' = -10$ cm. The magnification is $m = -(-10)/5 = 2$. The final image is *virtual*, *erect*, and *larger* than the object. This is the common way converging mirrors are used as "makeup mirrors" that present a magnified image of the user's face

when viewed from inside their focal length.

We only need to present *one* diagram for diverging/convex mirrors, as they all have the same general diagram independent of the relative size of s and f. Note that the first and second rules are "backwards" compared

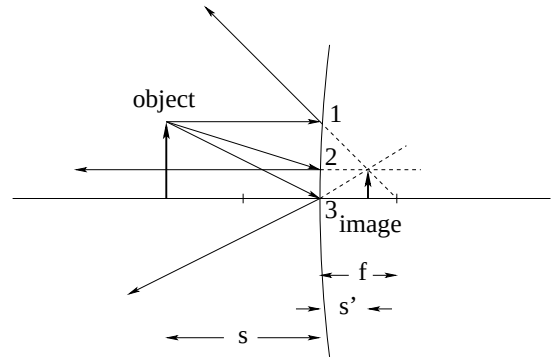


Figure 11.10: Converging mirror with $s = 20 < f = 10$.

to converging lenses. A ray parallel to the axis is deflected so it appears to be *coming from* the far side focal length. A ray headed *to* the far side focal length is deflected back parallel to the axis. The central ray is drawn as before.

We apply *as always* the mirror/thin lens formula: $1/s' = -1/10 - 1/20 = -3/20$ so $s' = -6.7$ cm. The magnification is $m = -(-6.67)/20 = 0.33$. The image is erect, virtual, and smaller than the object. All of these general properties will apply (with different numbers) to *any* diverging mirror.

If you master drawing these generic diagrams (and can manage the very simple algebra associated with evaluating e.g. s' and m given s and f, you can with patience analyze any combination of mirrors (and later) lenses) you are presented with.

11.4 Lenses

A spherical lensing surface between two different media with different indices of refraction are drawn in figure (11.11).

As was the case for the mirror, the three angles α, β, and γ in the small

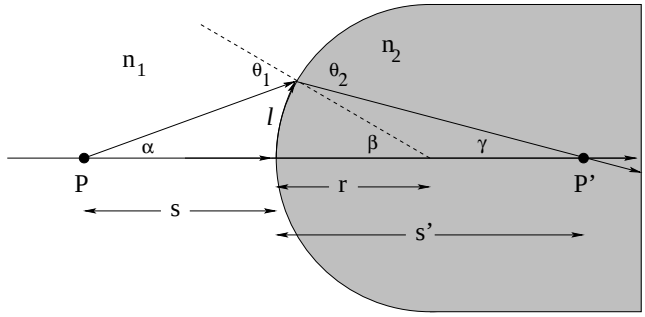


Figure 11.11: Diagram that shows how a spherical lens creates an image via refraction.

angle approximation can be written as:

$$\alpha \approx \frac{l}{s} \tag{11.25}$$

$$\beta = \frac{1}{r} \tag{11.26}$$

$$\gamma \approx \frac{1}{s'} \tag{11.27}$$

We also have *Snell's law* for the (small) angles θ_1 and θ_2:

$$n_1\theta_1 \approx n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \approx n_2\theta_2 \tag{11.28}$$

so

$$\theta_2 = \frac{n_1}{n_2}\theta_1. \tag{11.29}$$

Using triangle rules like the ones above, we also get:

$$\theta_1 = \alpha + \beta \tag{11.30}$$

and

$$\beta = \theta_2 + \gamma \tag{11.31}$$

Eliminating $theta_2$, this becomes:

$$\beta = \frac{n_1}{n_2}\theta_1 + \gamma \tag{11.32}$$

If we multiply both sides by n_2 and substitute θ_1 from the first equation, this becomes:

$$n_2\beta = n_1\alpha + n_1\beta + n_2\gamma \tag{11.33}$$

or

$$n_1\alpha + n_2\gamma = (n_2 - n_1)\beta \tag{11.34}$$

We substitute in the small angle formulas and cancel l to get:

$$\frac{n_1}{s} + \frac{n_2}{s'} = (n_2 - n_1)\frac{1}{r} \tag{11.35}$$

In most cases of interest to us, the lenses in question will be made out of glass, plastic, or collagen (in the case of the eye) surrounded or faced by air, in which case this will simplify to:

$$\frac{1}{s} + \frac{n}{s'} = (n-1)\frac{1}{r} \tag{11.36}$$

If there are two lensing surfaces separated by a very small distance, we have a so-called *thin lens*. The relevant geometry of a thin lens surrounded by air is shown in (11.12). The first surface struck by light from an object

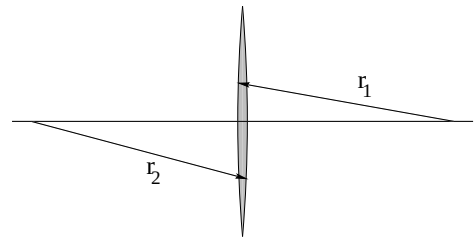


Figure 11.12: Geometry of a thin lens surrounded by air.

(presumed coming in from the left) has positive radius of curvature r_1. The second surface has a negative radius of curvature r_2. The index of refraction of the lens is n.

Suppose we have an object on the left hand side of this lens at distance s. From the formula above, we have:

$$\frac{1}{s} + \frac{n}{s'} = (n-1)\frac{1}{r_1} \tag{11.37}$$

The image of the first lensing surface is a *virtual object* for the second lensing surface. Because it is virtual (located to the *right* of the second surface, on the side light is going *to*) and because we are going from the material with index of refraction n into air, the formula for the second lensing surface is:

$$\frac{-n}{s'} + \frac{1}{s''} = (1-n)\frac{1}{r_2} \tag{11.38}$$

If we add these two formulae, the s' term cancels and, we get:

$$\frac{1}{s} + \frac{1}{s''} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = \frac{1}{f} \tag{11.39}$$

This is the *thin lens formula* where s'' is the final location of the image of the entire lens. Note that this is *identical* to the formula for the mirror. The focal length is given by the **lensmaker's formula**:

$$\frac{1}{f} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{11.40}$$

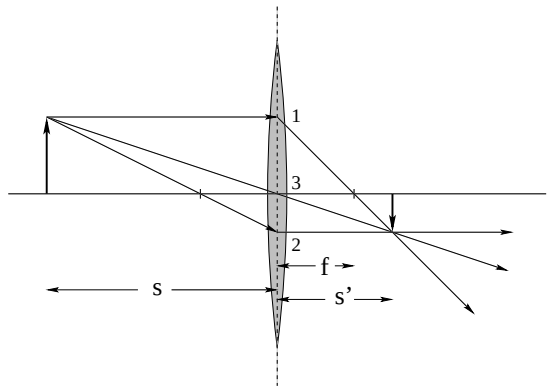


Figure 11.13: A converging lens with focal length of 10 cm and an object at $s = 30$ cm.

With the thin lens formula in hand, we can easily adapt *exactly* the same rules for drawing ray diagrams for locating images. Let's draw a simple ray diagram for a converging and a diverging lens that are similar to the ray diagrams above for mirrors. We do the usual algebra and arithmetic: $\frac{1}{s'} = \frac{1}{10} - \frac{1}{30} = \frac{2}{30}$ so $s' = 15.0$ cm, $m = -\frac{1}{2}$. The final image is inverted, real, and smaller than the object.

As before, if one puts an object inside the focal length it will make a magnified, erect, virtual image, if one exchanges the position of object and image in the example above, one will obtain an inverted, real image that is larger than the object.

A diverging lens, on the other hand, has only one generic diagram to be learned. It is basically the same as for the mirror, except that rays are transmitted through the thin lens (with all bending occurring at the thin

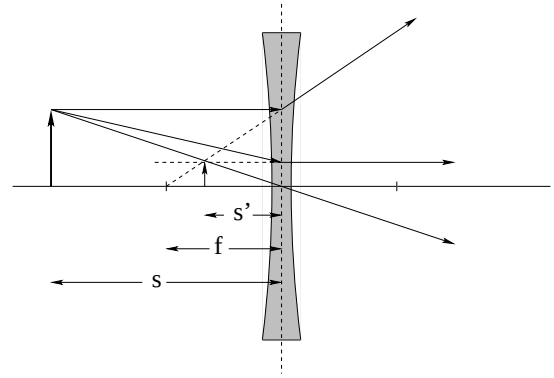


Figure 11.14: A diverging lens with focal length of -10 cm and an object at $s = 20$ cm.

plane representing the center plane of the lens) instead of reflected from it. In the situation represented in figure (11.14), the image is virtual, erect, and smaller than the original object. Show (from the numbers and thin lens formula) that $s' = -6.67$ cm and that $m = 1/3$.

11.5 The Eye

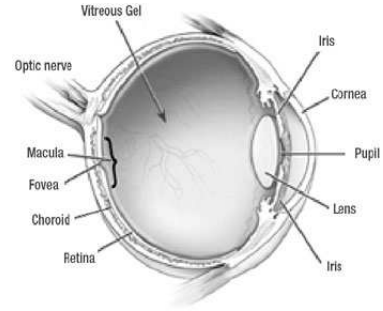


Figure 11.15: A simplified anatomical diagram of the human eye.

The eye is roughly spherical and approximately one inch in diameter. Figure (11.15) show is essential anatomy. Here is a brief review of the components of the eye.

- **Cornea:** The cornea of the eye is the rounded, transparent structure at the front of the eye. It is strongly curved, and is responsible for *most* of the bending of light required to focus images onto the...

- **Retina:** The retina is the "film" of the eye. It consists of tight bundles of photosensitive nerves called *rods* (sensitive to light intensity) and *cones* (sensitive to intensity in specific colors. In the center of the retina is the...

- **Macula:** The macula is the most sensitive part of the retina and is where one "sees" the object of one's attention. It is more or less in front of the...

- **Optic Nerve:** which pipes all of the information transduced from the light image cast on the retina to the brain. The retina (especially the macula) is very sensitive to light and easily damaged. To control the amount of light entering the eye, the...

- **Iris:** The iris is a ring of pigmented tissue that can open or contract to let more or less light into the...

- **Pupil:** The pupil is the aperture for light into the eye. When it is dark, the iris opens and lets all the light possible into the retina (which is very sensitive and capable of seeing with remarkably little light). When it is very bright, the iris closes down to a pinpoint. This actually increases visual acuity – see the *pinhole camera* – independent of the action of the...

- **Lens:** The lens of the eye is normally in a state of tension maintained by suspensory ligaments called **zonules** that keep it flattened out, with a maximally long focal length. A ring of **ciliary muscles** surrounding the lens can be contracted, which removes a part of this tension, predictably bulging the lens and thereby reducing its focal length. This process is called **accommodation**.

It is important to understand that accommodation can only *reduce* the focal length of the lens, not increase it, as well as the fact that the cornea is responsible for most of the focal length of the combined system – the actual lens is more of a "correction" to the overall focal length already achieved by the cornea alone. We now need to understand the three common conditions that describe the eye.

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

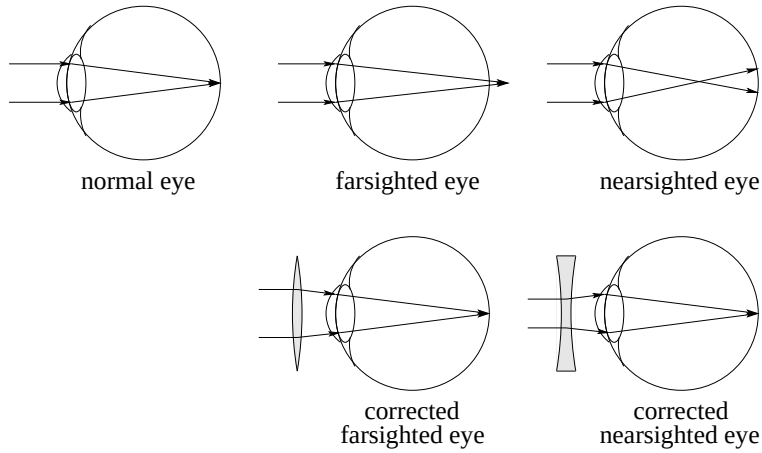


Figure 11.16: The focal length of the relaxed (combined) lensing acting of the eye for a normal eye, a farsighted eye (hyperopia), and a nearsighted eye (myopia).

automatically in focus (as a real image cast upon) on the retina. Given a distance from the cornea to the retina of roughly 2.5 cm, this means that the strength of the lens of a normal eye is approximately $\frac{1}{0.025} = 40.00$d. When viewing less distant objects, accomodation *shortens* the focal length to bring them into focus on the retina.

The focal length of a relaxed *farsighted* eye is *behind* the retina (too long, strength less than 40.00d) and is corrected with a *converging* lens to make up the difference. If one expresses strength in diopters, one can simply add a converging lens with a strength in diopters to the strength of the the eye to get the "right strength" to make the combination focus distant objects on the retina with the eye's lens relaxed. Note that a hyperopic person *can* see in focus all the way out to infinity, but they have to use accommodation to shorten their lens's "too long" relaxed focal length see even distant objects, which can lead to eye fatigue and headaches.

The focal length of a relaxed *nearsighted* eye is in *front* of the retina (too short, strength greater than 40.00d) and is corrected with a *diverging* lens to take *away* some of its strength. A myopic individual simply cannot see distant objects in focus without a corrective lens because accommodation cannot *increase* the focal length of the eye's lens, it can only further decrease it.

Accommodation can shorten the focal length only so far, which limits how close an object can be and still be focused on the retina. The nearest point one can bring an object to the eye and still bring it into focus on the retina is called the *near point* of the eye and is also the *distance of most distinct vision*, represented x_{np}. In most adults, this distance is around 25 cm (less for small children, longer for the elderly).

A nearsighted person's lens *already* has too short a focal length to be able to focus distant objects on the retina, and accommodation only shortens the focal length still farther. A nearsighted person cannot see anything clearly at distances *greater* than some point, called the *far point* for that person's eyes. A nearsighted person is one for whom the far point x_{fp} is less than infinity.

A common aberration of human eyes is a condition called **astigmatism**. Astigmatism is what happens when the eye's lens is no cylindrically symmetric. That is, the focal length of the lens in the horizontal plane is not the same as the focal length in the vertical plane. One can then bring things into focus in one dimension with accommodation, but only at the expense of blurring them in the other. The solution is to wear lenses that are astigmatic in the opposite direction to add up to neutral (or to person's otherwise necessary correction).

As a person's eyes age, their ability to focus changes. People with once normal vision can become nearsighted or farsighted. After the age of roughly 50 a new condition often emerges – that of **presbyopism**. The collagen of the lens hardens over time. Its flexibility decreases, making it more difficult for the eye to accommodate and *increasing the near point*. This kind of "farsightedness" can occur even for nearsighted individuals. The solution is to correct with "reading glasses" – positive lenses that permit a presbyopic individual to read at normal distances. They can be combined into "bifocals" – reading glasses for short distances plus diverging lenses to correct myopia at long distances – for people with the latter condition.

11.6 Optical Instruments

11.6.1 The Simple Magnifier

The "size" of an object to the human eye is determined by three distinct things. Humans have binocular vision, and use *parallax* – the apparent displacement of an object seen from two slightly different positions – to get a sense of an object's distance. This is reinforced by the physiological sense of *accommodation*, which gives one a sense of relative nearness. Finally, given the distance, it is determined by the *angle* the image subtends on the retina.

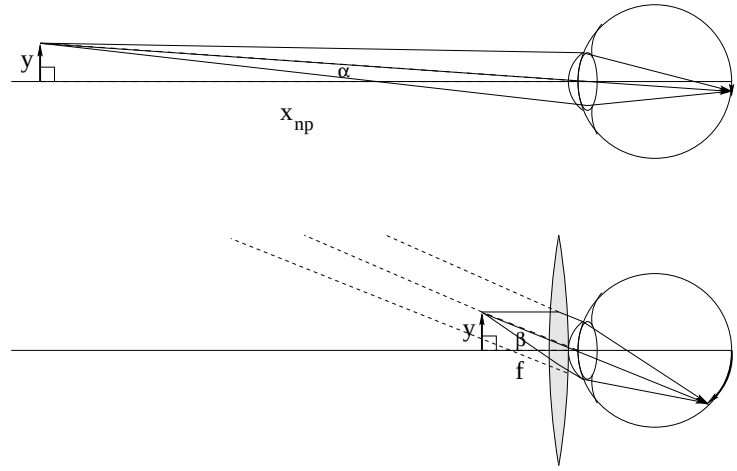


Figure 11.17: A converging lens used as a simple magnifier.

To see a small thing as clearly as possible, we naturally bring it to the closest point we can, so its details subtend the largest possible angle when our eyes are maximally accommodating. In figure (11.17) the top picture shows an object of height y viewed at the near point. When the image is focused on the retina by the maximally accomodating eye, it subtends an angle of α, where:

$$\alpha \approx \tan(\alpha) = \frac{y}{x_n p} \tag{11.41}$$

in the small angle approximation (which is entirely justified because we only "see" detail with the macula, which in turn only occupies around 0.2 radians in the center of the visual field. Even if we are examining a larger object,

we do so by redirecting the eye to look at it in patches that cover it in small angle chunks.

To use a simple magnifier we place a converging ($f > 0$) lens immediately in front of the eye. The object is placed at its focal point. It therefore forms a *virtual image* at $-\infty$ that is automatically brought into focus by the relaxed normal (or vision corrected) eye. It now subtends an angle β on the retina given by:

$$\beta \approx \tan(\beta) = \frac{y}{f} \tag{11.42}$$

The magnification is therefore the ratio of the new angle (with the magnifier) to the angle without it, when the object is seen at the near point. The magnification of the object occurs because one can bring the object *closer* to the eye than x_{np} and still see it clearly (more clearly, even, than before given that one does not have to accommodate). Its magnification is given by:

$$M = \frac{\beta}{\alpha} = \frac{x_{np}}{f} \tag{11.43}$$

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.

11.6.2 Telescope

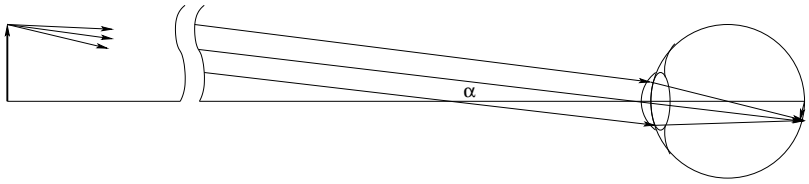


Figure 11.18: An regular (inverting) telescope.

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 (11.18) you can see what a ray diagram looks like for light from a very distant object entering the naked human eye. The rays from the originating point, after travelling a long distance, necessarily enter the eye more or less parallel and

are focused by the relaxed normal lens onto the single point on the retina determined by the central ray entering at angle α.

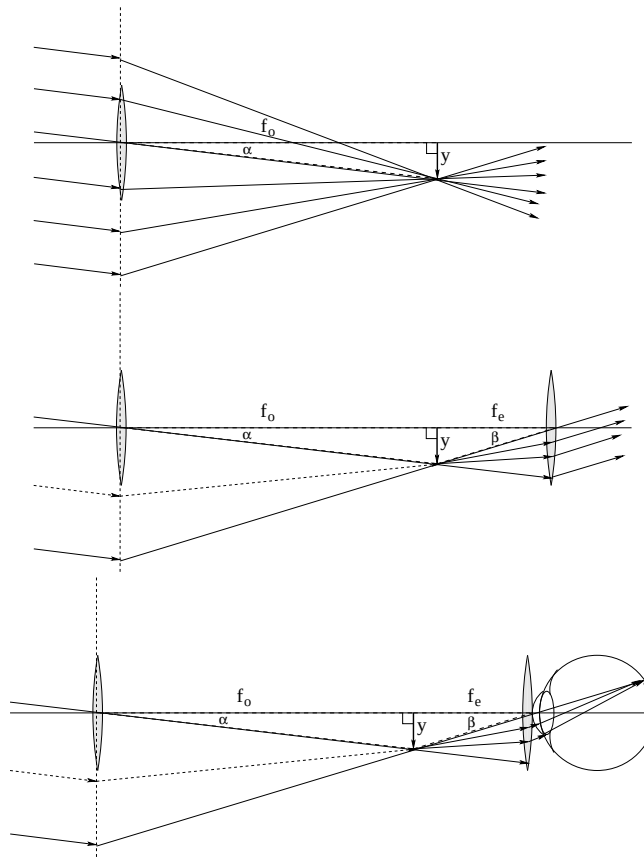


Figure 11.19: An regular (inverting) telescope.

To magnify our view of this object, we begin by inserting a lens with a *long* focal length f_o into the optical path. This takes light from the (infinitely) distant object and creates an *inverted real image of it* at the focal point as shown in the first panel in figure (11.19) 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 if much steeper. We can easily compute the angular magnification of this telescope by noting that:

$$\alpha \approx \tan(\alpha) = -\frac{y}{f_o} \tag{11.44}$$

and

$$\beta \approx \tan(\beta) = \frac{y}{f_e} \tag{11.45}$$

so that

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \tag{11.46}$$

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* lense for the eyepiece. Ordinarily one thinks of a diverging lens as making something smaller, but because we can place the image from the first lens anywhere we wish, we can turn it into a virtual object at the *far* focal point of a diverging lens. One obtains the same formula for the magnification, but now $f_e < 0$ and the overall angular magnification is *positive*.

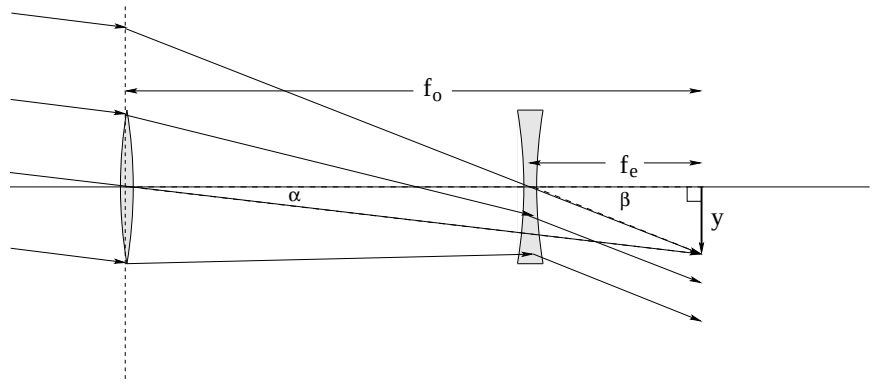


Figure 11.20: A "Galilean" telescope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the distant objects *erect* instead of inverted.

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 (11.20), the angular magnification of a Galilean telescope is still:

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \tag{11.47}$$

(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.

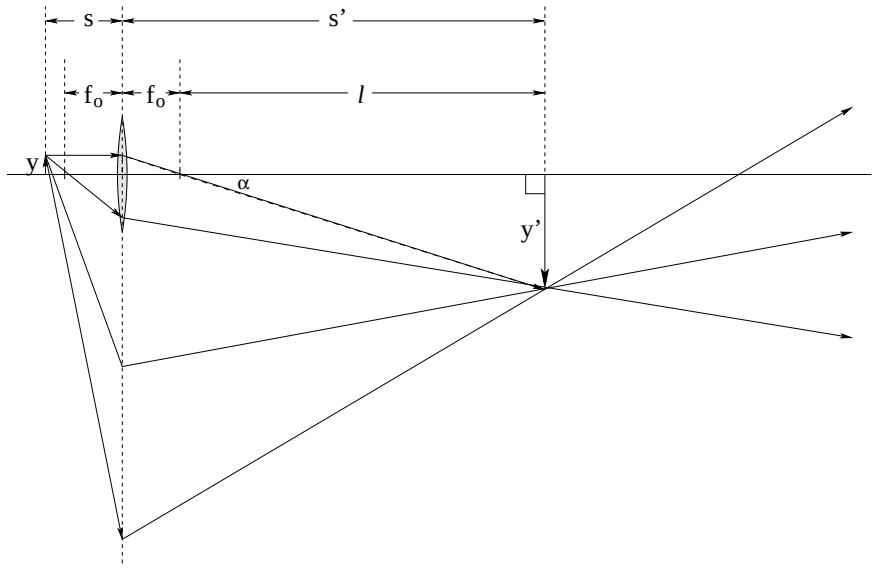


Figure 11.21: The first magnification stage of a compound microscope brings a *small* object just outside of the focal point of the objective lens into focus as a *real, magnified image* at the end of the **tube length** l. By comparing the two dashed similar triangles, one can see that the first stage magnification is $-\frac{l}{f_o}$.

11.6.3 Microscope

A compound microscope is used to view a very small, but nearby object. It accomplishes this in two stages. Two short focal length lenses are situated at ends of a tube much longer tube. The **tube length** l of the microscope is by definition the distance between the focal point of the first, or *objective* lens (which must be converging) and the second, or *eyepiece* lens.

The objective stage of the magnification occurs as the the object is placed on a movable platform just outside of the focal length of the objective lens of the microscope. The platform is raised or lowered (altering s, the object distance) until the objective lens forms a *magnified, real image* of the object at the end of the tube length as shown in figure (11.21).

The magnification of the objective stage is:

$$M_o = -\frac{\ell}{f_o} = -\frac{f_o + l}{s} \tag{11.48}$$

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.

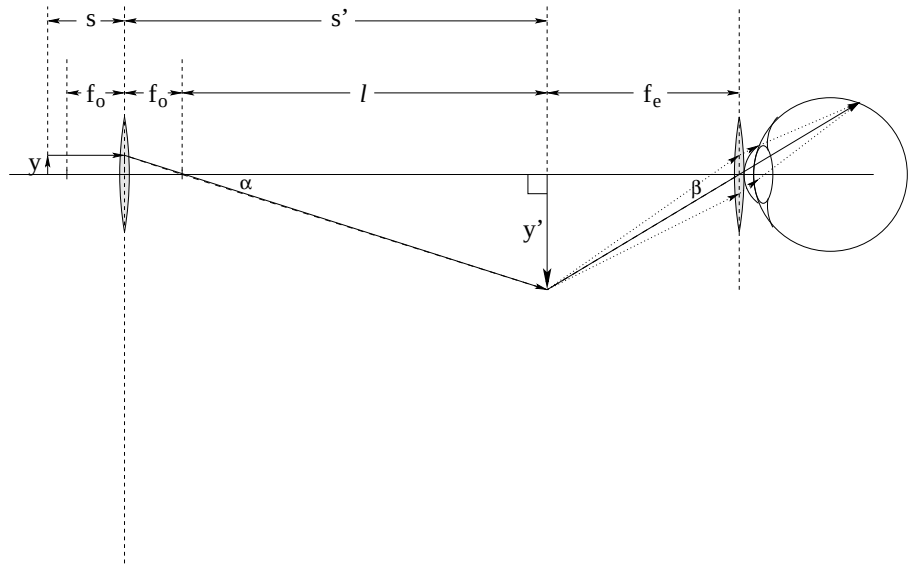


Figure 11.22: The second magnification stage of a compound microscope brings the *highly magnified* image from the objective stage close to the eye by functioning as a simple magnifier. By bringing the virtual image in from x_{np} to f_e it magnifies it by an addtional factor of $\frac{x_{np}}{f_e}$.

This real, magnified image can be viewed with the naked eye, but of course the naked eye can view it no *closer* than x_{np}. The second stage of a compound microscope consists of an eyepiece lens is used as a *simple magnifier* to view this real image in precisely the same way we used it for the telescope, and can be converging or diverging as was the case for the telescope. It produces a virtual image at infinity that subtends a greater angle than the real image formed by the objective lens alone would if viewed at the near point of the relaxed normal eye.

The magnification of the eyepiece used as a simple magnifier is therefore:

$$M_e = \frac{x_{np}}{f_e} \tag{11.49}$$

which yields an overall magnification for the two stages working together of:

$$M_{tot} = -\frac{\ell \; x_{np}}{f_o f_e} \tag{11.50}$$

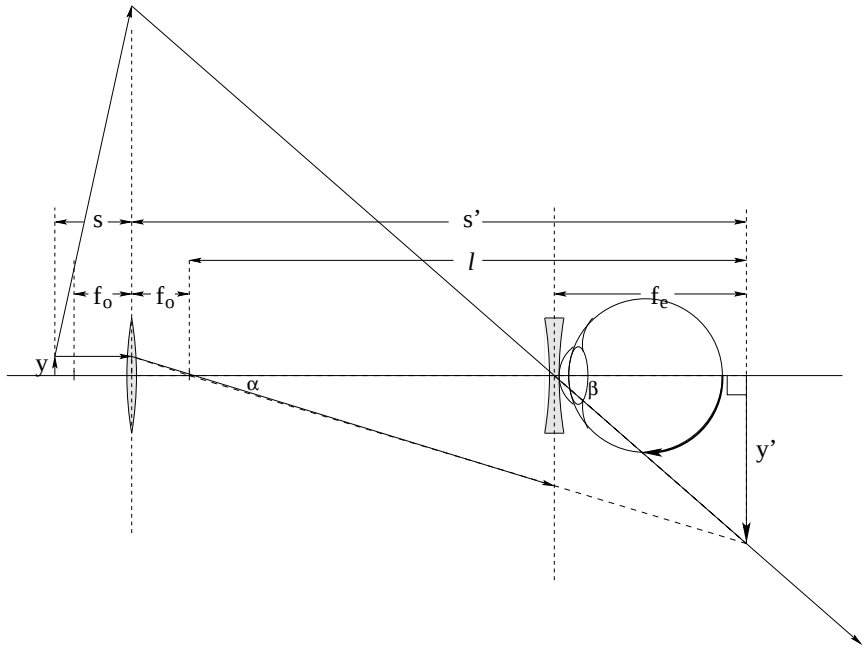


Figure 11.23: A "Galilean" microscope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the tiny objects *erect* instead of inverted. As always, we use a "central" ray for the second lens that is deflected at the plane of the first lens *as if* it passes through both lenses to find the location and size of the final image.

As we noted and can see in figure (11.23) above, one can use a diverging lens for the eyepiece by placing the real image formed by the objective on the *far* side of the diverging lens to form a **"Galilean" microscope**. As before (for the telescope) this microscope does not invert the image (inversion is inconvenient and undesireable) but otherwise the same formula works for the magnification provided that one uses a negative f_e for the diverging lens. It has the further advantage of having a slightly shorter overall length.

Typical numbers for a compound microscope this might be $f_o = f_e = 1$ cm, $l = 10$ cm, for a total magnification of 250 (inverting or non-inverting). 250x microscopes are *more than adequate* to observe e.g. blood cells, bacteria, the cellular structure of plant an animal tissue, amoeba, paramecium, and a host of microorganisms and cellular structures. For example, amoeba can range in size from 10-1000 μm (where the latter, note well, is roughly

a millimeter and barely visible to the naked eye). A 250 power microscope can make an amoeba appear to the eye as large as a 25 cm object, clearly revealing its nucleus and vacuoles. Even small amoeba or bacteria will appear several millimeters in size at this magnification.

Just as the telescope caused a revolution in our vision of cosmology and the structure of the Universe at large distances and over long times, the microscope caused a revolution in our vision of the world of biology. Disease, which had long been thought of as being caused by demons or by a curse afflicted on sinners by God, was seen to be caused by living organisms too small to be seen by the naked eye. Where before the only possible cure for most diseases was believed to be divine intervention, miracles brought about by repentance and prayer, the microscope enabled the discovery of antiseptic medicine – that heat, soap and water, alcohol, and eventually antibiotics kill off disease-causing microorganisms to prevent or cure disease quite independent of "magic" such as miracles or prayer. The two together brought about the Enlightenment, a time of intense discovery and invention that ultimately ushered in the rational modern world of today.

11.7 Homework for week 11

Problem 1.

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 local its point-image, then doing some work with triangles and the small angle approximation.

Problem 2.

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 3.

Prove that the *lateral magnification* or an object is:

$$m_l = \frac{\Delta s'}{\Delta s} = \frac{s'^2}{s^2} \tag{11.51}$$

I'd "suggest" that you think about your friend, the *binomial expansion*, when solving this problem. Is the image "inverted"?

Problem 4.

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 5.

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 6.

Draw a ray diagram for the simple magnifier, deriving its (angular) magnification in the standard picture. Then derive where one has to locate the object to form a virtual erect image at the near point of the eye as viewed through the magnifier. What is the overall (angular) magnification of the image now?

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.

From the first 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 accomodation 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?

Week 12: Interference and Diffraction

- **Coherence:** A wave is said to be **coherent** 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/length of a typical hot source (such as a light bulb) is a few tens of periods or wavelengths.
- The coherence length of a laser can be as long as meters.

12.1 Harmonic Waves and Superposition

Several weeks ago we learned about **harmonic waves**, solutions to the wave equation of the general form (in one dimension):

$$\boldsymbol{E}(x,t) = E_0\hat{\boldsymbol{\epsilon}}\sin(kx-\omega t) \tag{12.1}$$

where $\hat{\boldsymbol{\epsilon}}$ 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:

$$\boldsymbol{E}(r,t) = E_0\frac{a}{r}\hat{\boldsymbol{\epsilon}}\sin(kr-\omega t) \tag{12.2}$$

(where $|\boldsymbol{E}(a,t)| = E_0$ is the field strength at the surface of the source for this component of the polarization). Recall also that we only need to write the

electric field strength because the associated magnetic field has an amplitude of $B_0 = E_0/c$, is in phase, and is perpendicular to the electric field so that the Poynting vector:

$$\boldsymbol{S} = \frac{1}{\mu_0}\boldsymbol{E} \times \boldsymbol{B} \tag{12.3}$$

points in the direction of propagation. Finally, don't forget that the (time averaged) intensity of the wave is:

$$I_0 =< |\boldsymbol{S}| >_{\text{av}} = \frac{1}{2\mu_0} E_0 B_0 = \frac{1}{2\mu_0 c} E_0^2 \tag{12.4}$$

We also learned about **Huygen's principle**, which states that each point on a wavefront of a propagating harmonic wave acts like a *spherical source* for the future propagation of the wave. This will prove to be a key idea in understanding interference and diffraction of waves that pass through slits, the **superposition principle**, which says that to find the total field strength at a point in space produced by waves from several sources we simply add the field strengths from all the sources up, and one of the ideas underlying Snell's law, that the wavelength of a wave of a given fixed frequency depends on the index of refraction of the medium through which it propagates according to:

$$\lambda' = \frac{\lambda}{n} \tag{12.5}$$

where λ is the wavelength in free space; the wavelength of a wave is *shorter* in a medium with an index of refraction greater than 1 so that the wave slows down. All of these things that we have already learned will be important in our development of interference and diffraction.

In addition to these old concepts, we will require one or two new ones. One is the idea of a *hot source*. A hot source is something like the hot filament of a light bulb, the hot flame of a candle, the hot gasses on the surface of the sun, all *so* hot that they glow and give off light. Even the gasses in a relatively cool fluorescent tube are "hot" in the sense we wish to establish, as the atomes that are giving off the light are very weakly correlated with one another.

Hot sources have certain important properties. The ones that are important to us are:

- Atoms (or molecules) emitting light of any given wavelength in one part of the hot source are not correlated with atoms in other parts that are emitting light at the same wavelength. Even though they have a *common wavelength*, the light from different regions have *random phases* with respect to one another.

- Atoms (or molecules) emitting light of any given wavelength from *one* part of the hot source are not self-correlated for long times. After a certain amount of time, the source picks up a *random phase* relative to its phase at a previous time.

One can think of the atoms as being little charged resonant oscillators that give off light at particular frequencies (and thereby damp their own oscillation) as they bounce. The "hot" part means that thermal energy from the lattice constantly "kicks" the atoms to add back energy that is radiating away and thereby cooling the atoms, but it kicks them at *random times and places* and thereby introduces a random phase as it does so.

In order to observe intereference or diffraction, we will need (as we will see) for there to be a *fixed* phase difference between the light fields we add up from different sources. We call light with this sort of fixed, reliable phase difference (at a given frequency) **coherent light**. We can characterize the coherence of the source by either how *long* it remains coherent (**coherence time**) or by the length of a harmonic train that remains coherent (**coherence length**).

Hot sources are generally quite incoherent – they remain coherent for (typically) a few tens to hundreds of periods of the wave, or equivalently for tens to hundreds of wavelengths. This means that one will generally not be able to observe interference or diffraction in light being recombined from several hot sources (slits or surfaces of reflection) unless the path difference is *smaller than the coherence length* of the light. We can see swirling colors in a soap bubble (a very thin film) from thin film interference in the reflected light because the bubble is thinner than the coherence length of the hot-source light illuminating it. We do not see similar colors in light reflected from a drinking glass because the glass is much thicker than the coherence length of the light.

As we will later prove, when light from two *incoherent* sources is added, one simply adds the intensities to find the final intensity. The proof involves

adding the light from two sources with an arbitrary phase difference, then averaging over all possible phase differences. When light is added from two coherent sources, the field strengths are added (accounting for any *fixed* phase difference due to e.g. path differences) and the result squared to obtain the final additive intensity.

We have examples of sources of light that are *not* "hot sources" and that have *little* randomness in the phase of the emitters giving off the light. First and foremost of these is the *laser*, which is an extremely coherent source. Lasers have coherence lengths measured in meters, not microns. They are so coherent that two *different* laser sources will still interfere – even though the sources have a random phase in between them it is a *fixed* random phase.

Pay careful attention to coherence as you work through interference and diffraction below. Remember, hot sources will usually produce interference when the light being summed is within the mutual coherence time/length of the light source in question.

12.2 Interference from Two Narrow Slits

The first, and simplest, example of interference is monochromatic (constant wavelength) light falling upon two extremely narrow (slit width less than the wavelength of the light) separated by a distance d that is order of a few wavelengths in size. Because the slits are so close together, they are within the correlation length even of most (monochromatic) hot sources, so that two slit interference patterns can easily be produced.

To compute the interference pattern produced by two slits, we begin by examining figure (12.1), wherein light of fixed wavelength λ falls normally onto a blocking screen through which two narrow slits have been cut. Each slit is so narrow that it acts like a "point" Huygens radiator. Light from one slit (the upper) travels a long distance and falls on a distant screen. Light from the lower slit travels this distance plus the *additional* distance $d\sin(\theta)$ to arrive at the same point.

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.

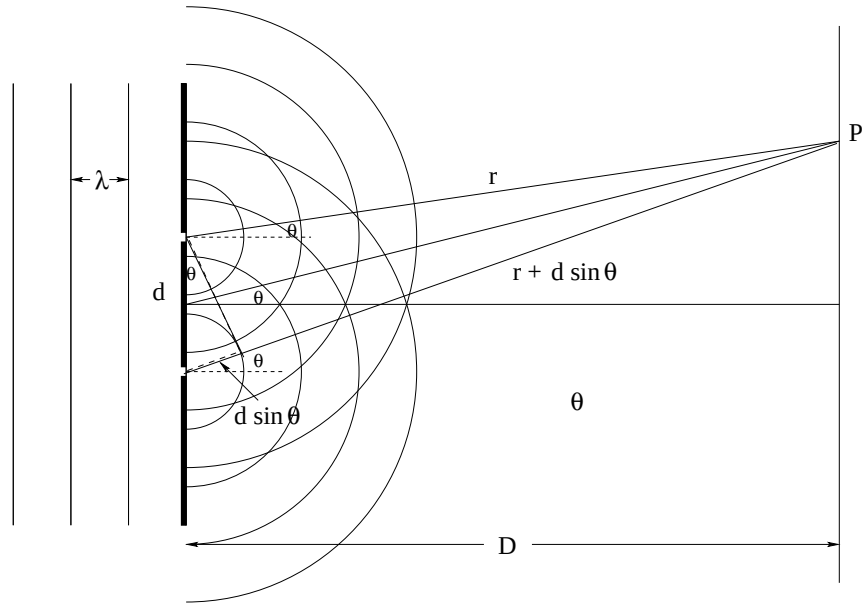


Figure 12.1: Two narrow slits act as Huygens radiators when indident plane wavefronts fall upon them. Light from the two slits is *coherent* and *in phase* as it leaves the slits, but arrives at P with a phase difference that depends on the path difference.

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 *at* point P E_0. Then the total field at point P is:

$$\begin{aligned} E_{\rm tot}(P) &= E_0\frac{a}{r}\sin(kr-\omega t) + E_0\frac{a}{r+\Delta r}\sin(kr+k\Delta r-\omega t) \\ &= E_0\frac{a}{r}\sin(kr-\omega t) + E_0\frac{a}{r}\left(1+\frac{\Delta r}{r}\right)^{-1}\sin(kr+k\Delta r-\omega t) \end{aligned}$$

$$
\begin{aligned}
&= E_0\frac{a}{r}\sin(kr-\omega t) + E_0\frac{a}{r}\left(1-\frac{\Delta r}{r}+...\right)\sin(kr+k\Delta r-\omega t) \\
&= E_0\frac{a}{r}\sin(kr-\omega t) + E_0\frac{a}{r}\sin(kr+k\Delta r-\omega t) + \mathcal{O}\left(\frac{\Delta r}{r}\right) \\
&\approx E_0\sin(kr-\omega t) + E_0\sin(kr-\omega t+\delta) \qquad (12.6)
\end{aligned}
$$

where in the end we set $a = r$ (so the field amplitude of a single slit is E_0 as defined above), we neglect terms of order $\Delta r/r$ in the last expression, and we introduce the *phase shift produced by the path difference*:

$$\delta = kd\sin(\theta) = \frac{2\pi d}{\lambda}\sin(\theta) \qquad (12.7)$$

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.

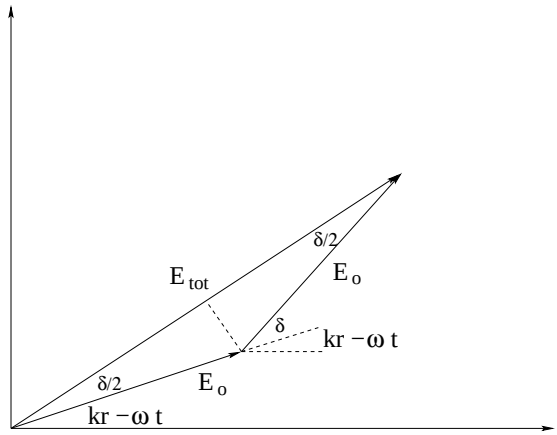


Figure 12.2: Phasor diagram for the addition of the electric field components of two slits.

In figure (12.2) we see the requisite phasor geometry. The light from the first slit has a field amplitude of the y-component of a "vector" (phasor) of

length E_0 at angle $kr - \omega t$ with respect to the x-axis. The light from the second slit is the y-component of a phasor of length E_0 at angle $kr - \omega t + \delta$. The field amplitude of the sum is the y-component of the phasor that is the vector sum of these two phasors, added by putting the tail of the second at the head of the first. Since the triangle representing this sum is isoceles it is easy to see that the two acute angles must both be $\delta/2$. The total amplitude is thus the sum of the adjacent side lengths of the two right triangles formed by dropping a normal as shown:

$$|E_{\rm tot}| = 2E_0 \cos(\delta/2) \tag{12.8}$$

and the full time dependent electric field is given by:

$$E_{\rm tot} = 2E_0 \cos(\delta/2) \sin(kr - \omega t + \delta/2) \tag{12.9}$$

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:

$$I_0 = \frac{1}{2\mu_0 c}|E_0|^2 \tag{12.10}$$

(from the Poynting vector, as we have seen many times at this point). The total intensity from the pair of slits is therefore:

$$I_{\rm tot} = 4I_0 \cos^2(\delta/2) \tag{12.11}$$

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:

$$\delta/2 = \pm\pi/2, \pm 3\pi/2, \pm 5\pi/2, ... \tag{12.12}$$

or

$$\delta = \frac{2\pi d}{\lambda}\sin(\theta) = \pm(2m+1)\pi \tag{12.13}$$

or the actual angles θ where:

$$d\sin(\theta) = \pm\frac{2m+1}{2}\lambda \tag{12.14}$$

The intensity is zero at the minima.

The maxima occur at the angles where:

$$\delta/2 = 0, \pm\pi, \pm 2\pi ... \tag{12.15}$$

or

$$\delta = \frac{2\pi d}{\lambda}\sin(\theta) = m2\pi \tag{12.16}$$

or the actual angles θ where:

$$d\sin(\theta) = \pm m\lambda \tag{12.17}$$

The intensity is $4I_0$ at the maxima.

12.3 Interference from Three Narrow Slits

In the case of three slits, each separated by the same distance d, we can follow a more or less identical procedure to find the overall amplitude from a phasor diagram. Consider the general phasor diagram above. We wish to add:

$$E_{\text{tot}} = E_0\sin(kr - \omega t) + E_0\sin(kr - \omega t + \delta) + E_0\sin(kr - \omega t + 2\delta) \tag{12.18}$$

with $\delta = kd\sin(\theta)$ is the phase angle produced by the path difference between any two adjacent slits. Examining figure (12.3) we see that the general

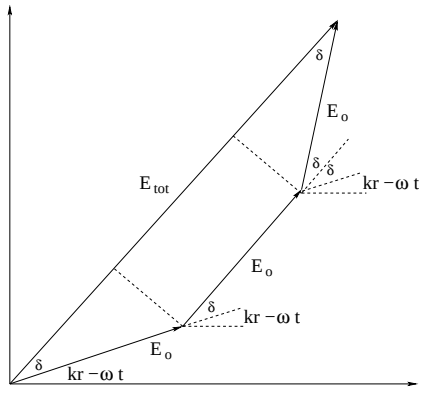


Figure 12.3: Phasor diagram for general solution for three slits.

result is:

$$E_{\rm tot} = E_0(1 + 2\cos(\delta)) \tag{12.19}$$

and we rather expect that the interference pattern intensity will be:

$$I_{\rm tot} = \frac{1}{2\mu_0 c}|E_{\rm tot}|^2 = I_0\left(1 + 4\cos(\delta) + 4\cos^2(\delta)\right) \tag{12.20}$$

which equals $9I_0$ when $\delta = 0, 2\pi, 4\pi...$ and equals I_0 when $\delta = \pi, 3\pi, 5\pi....$ 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 draw in figure (12.4):

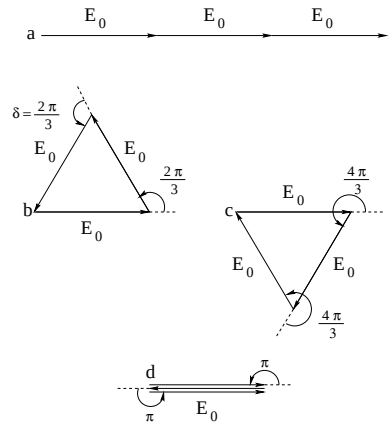


Figure 12.4: Phasor diagram illustrating (a) principle maxima; (b) first minimum; (c) second minimum; (d) secondary maximum.

The first arrangement in (a) shows three phasors lined 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 $\delta = 2\pi, 4\pi, 6\pi...$ – 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:

$$\frac{2\pi}{\lambda}d\sin(\theta) = \delta = 0, \pm 2\pi, \pm 4\pi... = \pm 2\pi\ m \qquad m = 0, 1, 2... \tag{12.21}$$

If we divide by 2π and multiply by λ, we see that this corresponds to:

$$d\sin(\theta) = \pm m\lambda \tag{12.22}$$

just as before for two slits separated by d. 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.

When we wish to find the minimal we note that the intensity is non-negative. The smallest it can possibly be is zero. 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 triangle. The two triangles starting at (b) and (c) in the figure above thus represent minima. We observe that we close the triangle when $\delta = \pm 2\pi/3, \pm 4\pi/3$, 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:

$$d\sin(\theta) = \frac{\pm m\lambda}{3} \qquad m =, 1, 2, , 4, 5, , 7, 8... \tag{12.23}$$

12.4 Homework for week 12

(Due 4/22/09)

Problem 1.

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 2.

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 $\theta \in [-\pi/2, \pi/2]$.

Problem 3.

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 $f = 0$ *and* where $\frac{df}{dx} = 0$ independently. Also recall that:

$$\lim x \to 0 \frac{\sin(x)}{x} = 0$$

(and is not undefined).

Problem 4.

Redo problem 1, 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 5.

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 6.

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 theta for $d = 4\lambda$, graph the intensity, and compare it to the answer to problem 1 above.

Problem 7.

Joe Braggart claims to have really, really good vision. "Why," he says. "My vision is *so* good I can make out the Galilean moons of Jupiter with my naked eyes on a really clear night. If I'd been around at the time of Galileo 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 *any chance* that he's telling the truth? (Note well: This is a problem on resolution, not lenses or the sensitivity of the retina.)

Problem 8.

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.

Week 13: Catchup and Conclusions

(Est 4/22-4/29)