

PH223: Physics for Chemists and Mechanical Engineers

R. Todd Lines
Brigham Young University-Idaho

Created in Scientific Workplace™

Copyright ©2011 by Author

Preface

Preface

This document contains my lecture notes for a new, experimental course. The goal of the course is to teach the introductory physics of waves, optics, and electricity and magnetism for mechanical engineering students.

Forward

Acknowledgments

BYU-I

R. Todd Lines.

Contents

Preface iii

Preface iii

Forward iii

Acknowledgments iii

1 Where We Start 1

What is this class? 1

Simple Harmonic Motion 3

SHM 3

Mathematical Representation of Simple Harmonic Motion 6

An example of oscillation 11

Energy of the Simple Harmonic Oscillator 18

Circular Motion and SHM 20

The Pendulum 23

Damped Oscillations 24

Driven Oscillations and Resonance 27

2 What is a Wave? 31

What is a Wave? 31

Wave speed 33

Example: Sound waves 34

One dimensional waves 36

3 Waves in One and More Dimensions 41

Sinusoidal Waves 41

The speed of Waves on Strings 51

Waves in two and three dimensions 52

4 Light, Sound, Power 57

Waves in matter-Sound 57

Speed of Sound Waves 61

Waves in fields-Light 62

Power and Intensity 63

5 Doppler Effect and Superposition 69

Doppler Effect 69

Superposition Principle 76

Superposition and Doppler: Shock waves 79

Importance of superposition 81

6 Standing Waves 83

Mathematical Description of Superposition 83

Reflection and Transmission 87

Mathematical description of standing waves 92

Standing Waves in a String Fixed at Both Ends 94

7 Light and Sound Standing waves 97

Sound Standing waves (music) 97

Lasers and standing waves 102

Standing Waves in Rods and Membranes	103
8 Single Frequency Interference, Multiple Dimensions	107
Mathematical treatment of single frequency interference	108
Single frequency interference in more than one dimension	116
9 Multiple Frequency Interference	123
Beats	123
Non Sinusoidal Waves	125
Frequency Uncertainty for Signals and Particles	137
10 Interference of light waves	141
The Nature of Light	141
Measurements of the Speed of Light	143
Interference and Young's Experiment	147
Double Slit Intensity Pattern	151
11 Many slits, and single slits	157
Diffraction Gratings	157
Single Slits	161
Narrow Slit Intensity Pattern	163
12 Apertures and Interferometers	169
Circular Apertures	169
Interferometers	172
Diffraction of X-rays by Crystals	175
Transition to the ray model	176
13 Ray Model	179

The Ray Approximation in Geometric Optics	179
Reflection	184
Reflections, Objects, and seeing	188
14 Refraction and images	191
Refraction	192
Total Internal Reflection	197
Images Formed by Refraction	201
15 Dispersion and Thin Lenses	205
Dispersion	205
Ray Diagrams for Lenses	210
16 Image Formation	221
Thin lenses and image equation	221
Thin Lenses	226
Images formed by Mirrors	232
Mirror reversal	234
17 Optical systems	243
Combinations of lenses	243
The Camera	246
18 Eyes and magnifiers	251
The Eye	251
Optical Systems that Magnify	255
Telescopes	260
19 Resolution and Charge	263

Resolution	263
Charge model	269
20 Electric charge	277
Charge	277
Insulators and Conductors	278
Conduction in solids	281
Note on drawing charge diagrams	287
21 Coulomb's Law and Lines of Force	289
Coulomb's Law	289
Direction of the force	293
More than two charges	294
Fields	300
Field Lines	304
On-Line demonstrations	307
22 Electric Fields of Standard Charge Configurations Part I	309
Standard Charge Configurations	309
Point Charges	309
Combinations of many charges	319
On-Line Visualizations	321
23 Electric Fields of Standard Charge Configurations Part II	323
Fields from Continuous Charge Distributions	323
24 Motion of Charged Particles in Electric Fields	333
Sheet of Charge	333

Constant electric fields	338
Particle motion in a uniform field	341
25 Dipole motion, Symmetry	351
Dipole motion in an electromagnetic field	351
Symmetry	359
26 Electric Flux	365
The Idea of Flux	365
27 Gauss' Law and its Applications	379
Gauss' Law	379
Examples of Gauss' Law	383
Derivation of Gauss' Law	390
28 Conductors in Equilibrium, Electric Potentials	395
Conductors in Equilibrium	395
Electrical Work and Energy	402
29 Electric potential Energy	407
Point charge potential energy	407
Dipole potential energy	415
Shooting α -particles	417
30 Electric Potentials	419
Example, potential of a capacitor	422
Electron Volt	426
31 Electric potential of charges and groups of charges	431

Point charge potential	431
Potential of groups of charges	439
32 Connecting potential and field	445
Finding the potential knowing the field	445
Sources of electric potential	450
Electrochemical separation of charge	451
batteries and emf	451
33 Calculating fields from potentials	459
Finding electric field from the potential	459
Geometry of field and potential	463
Conductors in equilibrium again	468
34 Capacitance	473
Capacitance and capacitors	473
Combinations of Capacitors	480
35 Dielectrics and Current	485
Energy stored in a capacitor	485
Dielectrics and capacitors	488
Induced Charge	490
Electric current	492
36 Current, Resistance, and Electric Fields	497
Current and resistance	497
Current density	503
Conservation of current	505

37 Ohm's law 509

 Conductivity and resistivity 509

 Power in resistors 518

 Magnetism 519

38 Magnetic Field 523

 Fundamental Concepts in the Lecture 523

 Discovery of Magnetic Field 523

39 Current loops 533

 Magnetic field of a current 533

 Magnetic dipoles 539

40 Ampere's law, and Forces on Charges 545

 Ampere's Law 545

 Magnetic Force on a moving charge 553

 Motion of a charged particle in a *B*-Field 555

 Hall Effect 562

41 Magnetic forces on wires 565

 Magnetic forces on Current-Carrying wires 565

 Torque on a Current Loop 567

42 Permanent Magnets, Induction 575

 Finally, why magnets work 575

 Back to the Earth 586

 Induced currents 587

43 Induction 591

Motional emf	591
Eddy Currents	594
Magnetic flux	595
44 Faraday and Lenz	601
Fundamental Concepts in the Lecture	602
Lenz	602
Faraday	603
Return to Lenz's law	604
Pulling a loop from a magnetic field.	605
45 Induced Fields	609
Generators	609
Transformers	616
Induced Electric Fields	618
Relationship between induced fields	621
Electromagnetic waves	623
46 Inductors	627
Self Inductance	627
Energy in a Magnetic Field	629
Return to Non-Conservative Fields	641
Magnetic Field Energy in Circuits	648
Mutual Induction	650
47 The Electromagnetic field	655
Relative motion and field theory	655
Field Laws	667

48 Field Equations and Waves in the Field 671

Displacement Current 671

Maxwell Equations 677

49 Waves in the Field 679

Wave equation for plane waves 684

Properties of EM waves 686

The Electromagnetic Spectrum 693

50 Polarization 697

Polarization of Light Waves 697

Retrospective 706

51 Summary of Right Hand Rules 709

PH121 or Dynamics Right Hand Rules 709

PH223 Right Hand Rules 711

52 Some Helpful Integrals 715

53 Table of Physical Constants 717

1 Where We Start

Fundamental Concepts

What is this class?

This class is designed to teach the physics of wave motion, electricity and magnetism, and optics. We have three major goals. One is to teach the physics that is not covered by Statics, Dynamics, and the Engineering Electronics Course. This physics can affect the mechanical systems you will design, build, or test, so knowing this physics is a very good thing. The second objective is to teach a different method of thinking about how things work. The third goal is to describe electrical and wave motion enough that the quantum nature of atoms and molecules make sense as our chemists take physical chemistry.

In engineering, the design parameters are often the goal. In physics, the physical relationship is the goal. For design engineers, both views are useful and important. The design is no good if the underlying principles preclude it from working!

As an example, I once worked on an optics project with a strong mechanical component. The system had scanning mechanisms that were fantastic mechanical devices. It was part of an aircraft and integrated into the aircraft system. But the optical system required two lasers that were separated in wavelength by only a few nanometers. The chief engineer knew how to build all the systems, but did not understand the physics that required the close wavelength spacing. He judged that the difficulty in building the device at that wavelength spacing outweighed any benefit, and he changed the specs to give two wavelengths that were fifty nanometers apart. Fifty nanometers is a pretty small tolerance. Surely it would be good enough! The resulting product did not work. For two years he tried to fine tune the scanners, and servos to make it work. After ten million dollars and two years, he finally moved the wavelengths closer. The cost of the change was an extra \$100,000 dollars, about 1/100 of the cost of the mistake. The

2 Chapter 1 Where We Start

system worked, but since this was a race to market, the time lost and the reputation lost on the faulty product destroyed the viability of the business. It is a bad day when you and your friends lose your jobs because you made a fundamental physics mistake!

Physics courses stress how we know what we know. They support the discipline called *system engineering*, which deals with the design of new and innovative products. As a more positive example, the National Weather service often releases requests for proposed weather sensing equipment. Their request might say something like the following:

Measure the moisture of the soil globally from an altitude of 800 km with an accuracy of 5%. The suggested instrument is a passive microwave radiometer.

The job of a system engineer is to determine what type of instrument to build. What is the underlying principle that it will use to do its job? What signal processing will it need? What mechanical and electrical systems will support this? This must all be determined before the bearings and slip-rings, and structures can be designed and built.

The radiometer design that came out of this project is flying today (or one very like it based on the original design) and is a major part of the predictive models that tell us what the weather will be in a few days.

Because this type of reasoning is our goal, we will not only do typical homework problems, but we will also work on our conceptual understanding.

I will also emphasize a problem solving method that I used with my engineering team in industry. It is a structured approach to finding a solution that emphasizes understanding as well as providing a numeric answer for a particular design. When you are part of an innovative design team, you will have to repeat a calculation over and over again each time some other part of the sign changes. If you have produced a symbolic solution, a numerical model, or at least a curve, you are ready for any changes in specifications. But if you have just “found the answer” you will have to find that answer again every time the overall design specs change. This approach is too slow, and, at least in my team, would have you finding a new job because our design efforts were always done against exacting schedules and budgets. By thinking in a structured method, with an eye toward symbolic answers or relationships rather than end numbers, you will learn to be a more valuable engineer. The process we will use is the same approach I used to teach my new engineers in the defense industry. It has been proven useful over and over for decades.

This same problem solving process is useful in chemistry, particularly as you study physical chemistry.

So let's get started. To understand waves, we need to get the waves moving. You studied Oscillation in Dynamics or PH121. Oscillating systems are often the disturbance that starts a wave. We will begin with a review of oscillation.

Simple Harmonic Motion

You are, no doubt, an expert in simple harmonic motion (SHM) after your PH121 or Dynamics class. But this will get us warmed up for the semester. In class we will use our clickers and go through a few questions. We will usually use the clicker system to answer a few questions to test your understanding of the reading material. This allows me to not waste time on things you already know, and to help me find the ones you don't. Most lectures will consist of me asking you if you have questions, and then if you don't, I will ask you "clicker questions." Where there is reason to believe you don't understand (with a normal cutoff of 80% of the class answering correctly being our definition of "understanding"), I will use the material from these written lectures to teach the concepts. So we won't always go through all the ideas and skills demonstrated in these written lectures. If you feel you would have liked more explanation on something but we did not cover that concept in class because most people were "getting it," you can come and see me in my office.

SHM

Let's consider a mass attached to a spring resting on a frictionless surface¹. This mass-spring system can oscillate.

Question 223.1.1

Question 223.1.2

Question 223.1.3

Question 223.1.4

Question 223.1.5

Question 223.1.6

Question 15.3.8

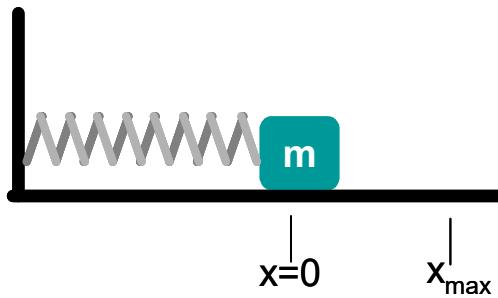
Question 15.3.9.2

Question 15.3.9.3

Question 15.3.9.5

Question 15.4

¹ Yes, I know there are no actual frictionless surfaces, but we are starting out at freshman level physics, so we will make the math simple enough that a freshman could do it by making simplifying assumptions. In this case, that the surface is frictionless.



Definition 1.1 Equilibrium Position: The position of the mass when the spring is neither stretched nor compressed.

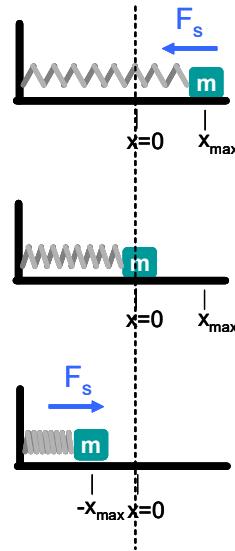
Hooke's Law

A law in physics is a mathematical expression of a mental model of how the universe works. Long ago it was noticed that the pull of a spring grew in strength as the spring was pulled out of equilibrium. The mathematical expression of this is

$$F_s = -kx \quad (1.1)$$

The force, F_s is directly proportional to the displacement from equilibrium, x . Since a man named Hooke wrote this down, it is called Hooke's law.

Hooke's Law is, strictly speaking, not a law that is always obeyed. It is a good model for most springs as long as we don't stretch them too far. We will often use the word "law" to mean *an equation that gives a basic relationship*. In that sense, Hook's law is a law.



Lets write Hooke's Law using Newton's second Law

$$\sum F_x = ma_x$$

If we assume no friction, we have just

$$-kx = ma_x$$

We can write this as

$$a_x = -\frac{k}{m}x \quad (1.2)$$

This expression says the acceleration is directly proportional to the position, and opposite the direction of the displacement from equilibrium. We can see that the spring force tries to oppose the change in displacement. We call such a force a *restoring force*.

Definition 1.2 *Restoring force: A force that is always directed toward the equilibrium position*

This is a good definition of *simple harmonic motion*.

Mathematical Representation of Simple Harmonic Motion

Recall from your Dynamics or PH121 classes that acceleration is the second derivative of position

$$a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$$

Hook's Law tells us

$$\begin{aligned} F &= ma = -kx \\ m \frac{d^2x}{dt^2} &= -kx \end{aligned}$$

We have a new kind of equation. If you are taking this freshman physics class as a... well... freshman, you may not have seen this kind of equation before. It is called a differential equation. But really the chances are that you are a sophomore or junior (or even a senior) and have lot of experience with differential equations. The solution of this equation is a function or functions that will describe the motion of our mass-spring system as a function of time. We will need to know this function, so let's see how we can find it.

Start by defining a quantity ω as

$$\omega^2 = \frac{k}{m} \quad (1.3)$$

why define ω^2 ? Because experience has shown that it is useful to define ω this way! But you probably remember ω as having to do with rotational speed, and from trigonometry (trig) you may remember using ω to mean angular frequency

$$\omega = 2\pi f$$

so our definition of ω may hint that k/m will have something to do with the frequency of oscillation of the mass-spring system.

We can write our differential equation as

$$\frac{d^2x}{dt^2} = -\omega^2 x \quad (1.4)$$

To solve this differential equation we need a function who's second derivative is the negative of itself. We know a few of these

$$\begin{aligned} x(t) &= A \cos(\omega t + \phi_o) \\ x(t) &= A \sin(\omega t + \phi_o) \end{aligned} \quad (1.5)$$

where A , ω , and ϕ_o are constants that we must find. Let's choose the cosine function

and explicitly take its derivatives.

$$\begin{aligned}x(t) &= A \cos(\omega t + \phi_o) \\ \frac{dx(t)}{dt} &= -\omega A \sin(\omega t + \phi_o) \\ \frac{d^2x(t)}{dt^2} &= -\omega^2 A \cos(\omega t + \phi_o)\end{aligned}$$

Let's substitute these expressions into our differential equation for the motion

$$\begin{aligned}\frac{d^2x}{dt^2} &= -\omega^2 x \\ -\omega^2 A \cos(\omega t + \phi_o) &= -\omega^2 A \cos(\omega t + \phi_o)\end{aligned}$$

As long as the constant ω^2 is our $\omega^2 = k/m$ we have a solution (now you know why we defined it as ω^2 !). Since from trig we remember ω as the angular frequency.

$$\omega = 2\pi f$$

Thus

$$\omega = \sqrt{\frac{k}{m}} = 2\pi f \quad (1.6)$$

The frequency of oscillation depends on the mass and the stiffness of the spring.

$$f = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \quad (1.7)$$

Let's see if this is reasonable. Imagine driving along in your student car (say, a 1972 Gremlin). You go over a bump, and the car oscillates. Your car is a mass, and your shock absorbers are springs. You have an oscillation. But suppose you load your car with everyone in your apartment². Now as you hit the bump the car oscillates at a different frequency, a lower frequency. That is what our frequency equation tells us. Note also that if we changed to a different set of shocks, the k would change, and we would get a different frequency.

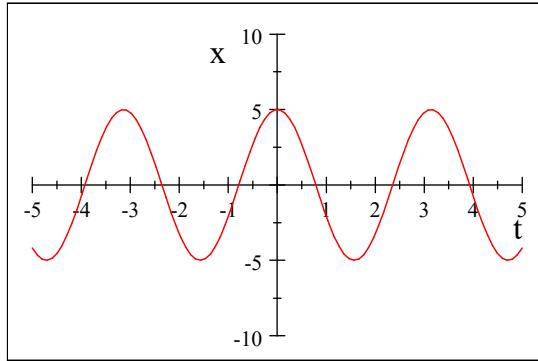
We still don't have a complete solution to our differential equation, because we don't know A and ϕ_o . From trigonometry, we recognize ϕ_o as the initial phase angle. We will call it the *phase constant* in this class. We will have to find this by knowing the initial conditions of the motion. We will do this in a minute.

A is the amplitude. We can find its value when the motion has reached its maximum displacement. Let's look at a specific case

$$\begin{aligned}A &= 5 \\ \phi_o &= 0 \\ \omega &= 2\end{aligned}$$

² If you are married, imagine taking two other couples with you in your car.

8 Chapter 1 Where We Start



We can easily see that the amplitude A corresponds to the maximum displacement x_{\max} .

Other useful quantities we can identify

We know from trigonometry that a cosine function has a period T .

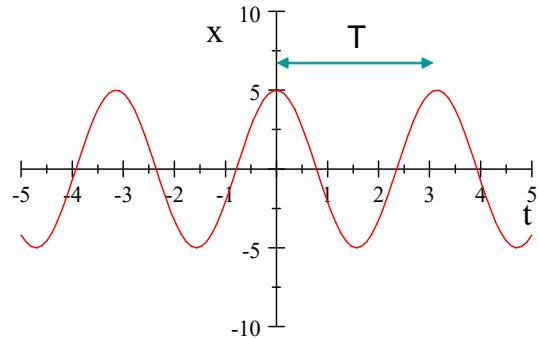


Figure 1.1.

The period is related to the frequency

$$T = \frac{1}{f} = \frac{2\pi}{\omega} \quad (1.8)$$

We can write the period and frequency in terms of our mass and spring constant

$$T = 2\pi\sqrt{\frac{m}{k}} \quad (1.9)$$

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \quad (1.10)$$

Velocity and Acceleration

Since we know the derivatives of

$$x(t) = A \cos(\omega t + \phi_o) \quad (1.11)$$

we can identify the velocity of the mass and its acceleration

$$v(t) = \frac{dx(t)}{dt} = -\omega A \sin(\omega t + \phi_o)$$

Recall that $A = x_{\max}$

$$v(t) = \frac{dx(t)}{dt} = -\omega x_{\max} \sin(\omega t + \phi_o) \quad (1.12)$$

We identify

$$v_{\max} = \omega x_{\max} = x_{\max} \sqrt{\frac{k}{m}} \quad (1.13)$$

Likewise for the acceleration

$$a(t) = \frac{dv(t)}{dt} \quad (1.14)$$

$$= \frac{d}{dt}(-\omega x_{\max} \sin(\omega t + \phi_o)) \quad (1.15)$$

$$= -\omega^2 x_{\max} \cos(\omega t + \phi_o)$$

where we can identify

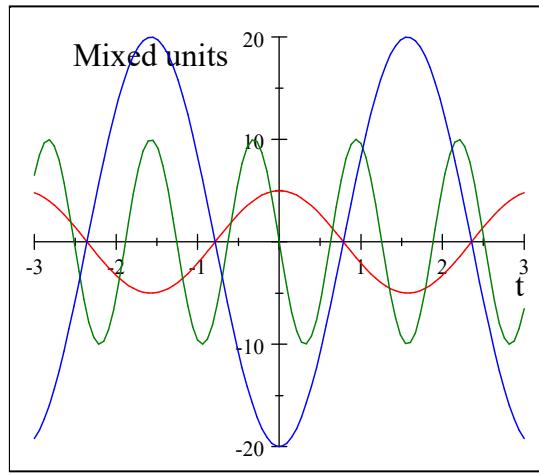
$$a_{\max} = \omega^2 x_{\max} = \frac{k}{m} x_{\max} \quad (1.16)$$

Comparison of position, velocity, acceleration

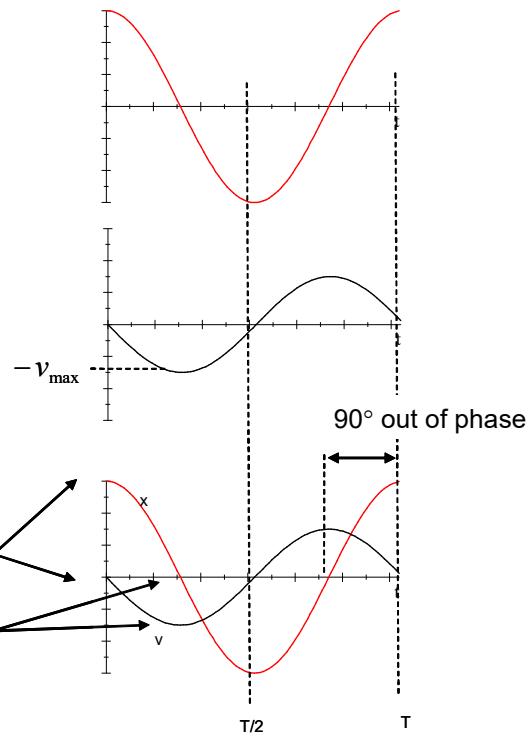
Don't do in class

Let's plot $x(t)$, $v(t)$, and $a(t)$ for a specific case

10 Chapter 1 Where We Start



Red is the displacement, green is the velocity, and blue is the acceleration. Note that each has a different maximum amplitude. That is a bit confusing until we recognize that they each have different units. We have just plotted them on the same graph to make it easy to compare their phases. Note that they are not in phase!



The acceleration is 90° out of phase from the velocity.

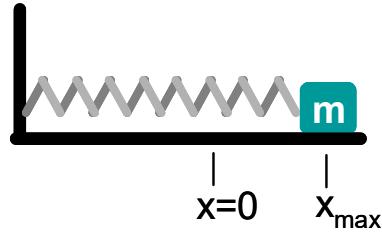
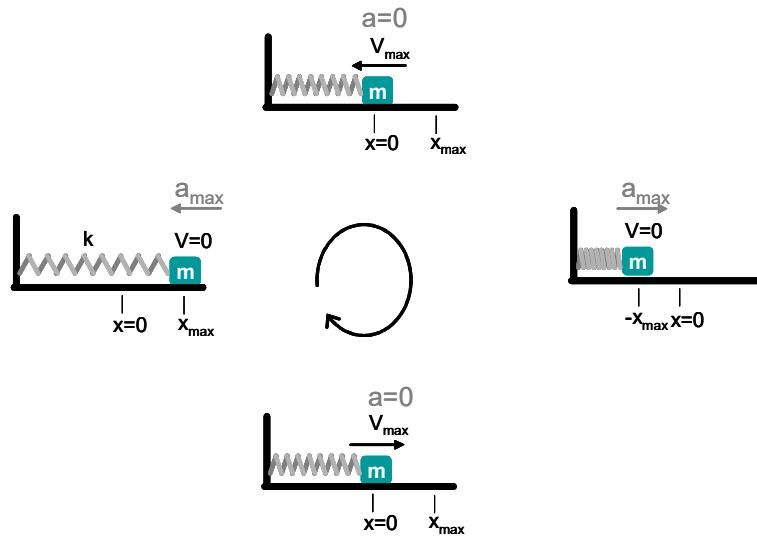


Figure 1.2.



An example of oscillation

We want to see how to find A , ω , and especially ϕ_0 . These quantities will be important in our study of waves. So let's do a problem.

Let's take as our system a horizontal mass-spring system where the mass is on a frictionless surface.

Initial Conditions

Now let's find A and ϕ_0 . To do this we need to know how we started the mass-spring motion. We call the information on how the system starts it's motion the *initial*

12 Chapter 1 Where We Start

conditions.

Suppose we start the motion by pulling the mass to $x = x_{\max}$ and releasing it at $t = 0$. These are our initial conditions. Let's see if we can find the phase. Our initial conditions require

$$\begin{aligned}x(0) &= x_{\max} \\v(0) &= 0\end{aligned}\tag{1.17}$$

Using our formula for $x(t)$ and $v(t)$ we have

$$\begin{aligned}x(0) &= x_{\max} = x_{\max} \cos(0 + \phi_o) \\v(0) &= 0 = -v_{\max} \sin(0 + \phi_o)\end{aligned}\tag{1.18}$$

From the first equation we get

$$1 = \cos(\phi_o)$$

which is true if

$$\phi_o = 0, 2\pi, 4\pi, \dots$$

from the second equation we have

$$0 = \sin \phi_o$$

which is true for

$$\phi_o = 0, \pi, 2\pi, \dots$$

If we choose $\phi_o = 0$, these conditions are met. Of course we could choose $\phi_o = 2\pi$, or $\phi_o = 4\pi$, but we will follow the rule to take the smallest value for ϕ_o that meets the initial conditions.

A second example

Using the same equipment, let's start with

$$\begin{aligned}x(0) &= 0 \\v(0) &= +v_i\end{aligned}\tag{1.19}$$

that is, the mass is moving, and we start watching just as it passes the equilibrium point.

$$\begin{aligned}x(0) &= 0 = x_{\max} \cos(0 + \phi_o) \\v(0) &= v_i = -v_{\max} \sin(0 + \phi_o)\end{aligned}\tag{1.20}$$

from

$$0 = x_{\max} \cos(\phi_o)$$

(first equation above) we see that³

$$\phi_o = \pm \frac{\pi}{2}$$

but we don't know the sign. Using our initial velocity condition

$$\begin{aligned} v_i &= -v_{\max} \sin\left(\pm \frac{\pi}{2}\right) \\ v_i &= -\omega x_{\max} \sin\left(\pm \frac{\pi}{2}\right) \end{aligned}$$

We defined the initial velocity as positive, and we insist on having positive amplitudes, so x_{\max} is positive. Thus we need a minus sign from $\sin(\phi_o)$ to make v_i positive. This tells us to choose

$$\phi_o = -\frac{\pi}{2}$$

with a minus sign.

Our solutions are

$$\begin{aligned} x(t) &= \frac{v_i}{\omega} \cos\left(\omega t - \frac{\pi}{2}\right) \\ v(t) &= v_i \sin\left(\omega t - \frac{\pi}{2}\right) \end{aligned}$$

Remark 1.1 Generally to have a complete solution to a differential equation, you must find all the constants (like A and ϕ_o) based on the initial conditions.

A third example

So far we have concentrated on finding ϕ_o . Let's do a more complete example where we find ϕ_o , A , and ω .

A particle moving along the x axis in simple harmonic motion starts from its equilibrium position, the origin, at $t = 0$ and moves to the right. The amplitude of its motion is 4.00 cm, and the frequency is 1.50 Hz.

a) show that the position of the particle is given by

$$x = (4.00 \text{ cm}) \sin(3.00\pi t)$$

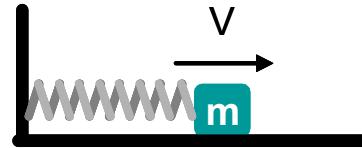
determine

b) the maximum speed and the earliest time ($t > 0$) at which the particle has this speed,

³ Really there are more possibilities, but we are taking the smallest value for ϕ_o as we discussed above.

14 Chapter 1 Where We Start

- c) the maximum acceleration and the earliest time ($t > 0$) at which the particle has this acceleration, and
- d) the total distance traveled between $t = 0$ and $t = 1.00\text{ s}$



Type of problem

We can recognize this as an oscillation problem. This leads us to a set of basic equations

Basic Equations

$$\begin{aligned}x(t) &= A \cos(\omega t + \phi_0) \\v(t) &= -\omega x_{\max} \sin(\omega t + \phi_0) \\a(t) &= -\omega^2 A \cos(\omega t + \phi_0)\end{aligned}$$

$$\omega = 2\pi f$$

$$v_m = \omega x_m$$

$$a_m = \omega^2 x_m$$

$$T = \frac{1}{f}$$

We should write down what we know and give a set of variables

Variables

t	time, initial time =0	$t_o = 0$
x	Position, Initial position =0	$x(0) = 0$
v		
a		
x_m	x amplitude	$x_m = 4.00 \text{ cm}$
v_m	v amplitude	
a_m	a amplitude	
ω	angular frequency	
ϕ_o	phase	
f	frequency	$f = 1.50 \text{ Hz}$

Now we are ready to start solving the problem. We do this with algebraic symbols first

Symbolic Solution

Part (a)

We can start by recognizing that we can find ω because we know the frequency. We just use the basic equation.

$$\omega = 2\pi f$$

We also know the amplitude $A = x_{\max}$ which is given. Knowing that

$$x(0) = 0 = A \cos(0 + \phi_o)$$

we can guess that

$$\phi_o = \pm \frac{\pi}{2}$$

Using

$$v(0) = -\omega x_{\max} \sin\left(0 \pm \frac{\pi}{2}\right)$$

again and demanding that amplitudes be positive values, and noting that at $t = 0$ the velocity is positive from the initial conditions:

$$\phi_o = -\frac{\pi}{2}$$

We also note from trigonometry that

$$x(t) = x_{\max} \cos\left(2\pi ft - \frac{\pi}{2}\right)$$

which is a perfectly good answer. However, if we remember our trig, we could write this using

$$\cos\left(\theta - \frac{\pi}{2}\right) = \sin(\theta)$$

Then we have

$$\begin{aligned} x(t) &= x_{\max} \cos\left(2\pi ft - \frac{\pi}{2}\right) \\ &= x_{\max} \sin(2\pi ft) \end{aligned}$$

Part (b)

We have a basic equation for v_{\max}

$$\begin{aligned} v_m &= \omega x_{\max} \\ &= 2\pi f x_{\max} \end{aligned}$$

to find when this happens, take

$$v(t) = v_{\max} = -\omega x_{\max} \sin\left(2\pi ft - \frac{\pi}{2}\right)$$

and recognize that $\sin(\theta) = 1$ is at a maximum when $\theta = \pi/2$ so the entire argument of the sine function must be $\pi/2$ when we are at the maximum displacement, so

$$\frac{\pi}{2} = \left(2\pi ft - \frac{\pi}{2}\right)$$

or

$$\pi = 2\pi ft$$

then the time is

$$\frac{1}{2f} = t$$

Part (c)

Like with the velocity we must use a basic formula, this time

$$a(t) = -\omega^2 A \cos(\omega t + \phi_o)$$

but recognize that the maximum is achieved when $\cos(\omega t + \phi_o) = 1$ or when $\omega t + \phi_o = 0$

$$\begin{aligned} t &= \frac{\phi_o}{\omega} \\ &= \frac{-\frac{\pi}{2}}{2\pi f} \\ &= \frac{-1}{4f} \end{aligned}$$

The formula for a_{\max} is

$$\begin{aligned} a_{\max} &= -\omega^2 x_{\max} \\ &= -(2\pi f)^2 x_m \end{aligned}$$

Part (d)

We know the period is

$$T = \frac{1}{f}$$

We should find the number of periods in $t = 1.00 \text{ s}$

$$N_{\text{periods}} = \frac{t}{T}$$

and find the distance traveled in one period, and multiply them together. In one period the distance traveled is

$$d = 4x_m$$

$$d_{\text{tot}} = d * \frac{t}{T} = 4fx_m t$$

Numerical Solutions

We found algebraic answers (or symbolic answers) to the parts of our problem above. We will always do this first. Then substitute in the numbers to find numeric answers.

Part (a)

$$\begin{aligned} x(t) &= x_{\max} \sin(2\pi ft) \\ &= (4.00 \text{ cm}) \sin(3.00\pi t) \end{aligned}$$

Part (b)

$$\begin{aligned} v_m &= 2\pi (1.50 \text{ Hz}) (4.00 \text{ cm}) \\ &= 0.377 \frac{\text{m}}{\text{s}} \end{aligned}$$

$$\begin{aligned} \frac{1}{2f} &= t \\ \frac{1}{2(1.50 \text{ Hz})} &= t \\ &= 0.333 \text{ s} \end{aligned}$$

Part (c)

$$\begin{aligned} t &= \frac{-1}{4f} \\ &= -0.16667 \text{ s} \end{aligned}$$

$$\begin{aligned}
 a_{\max} &= (2\pi f)^2 x_m \\
 &= (2\pi 1.5 \text{ Hz})^2 (4.00 \text{ cm}) \\
 &= 3.5531 \frac{\text{m}}{\text{s}^2}
 \end{aligned}$$

Part (d)

$$\begin{aligned}
 d_{tot} &= 4fx_m t \\
 &= 4 \times 4.00 \text{ cm} * 1.50 \text{ Hz} * 1.00 \text{ s} \\
 &= 0.24 \text{ m}
 \end{aligned}$$

We should make sure the units check. We put in units along the way, so we can be confident that they do. But if you did not work along the way with units, check them now.

We should also make sure our answers are reasonable. If the amplitude came out to be a billion miles, you might guess something went wrong. Always look over your answers to make sure they seem reasonable.

Energy of the Simple Harmonic Oscillator

Stop class here

If there is motion, there is energy. We can find the energy in a harmonic oscillator. Let's start with kinetic energy. Recall that

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

for our Simple Harmonic Oscillator (SHO) we have

$$\begin{aligned}
 K &= \frac{1}{2}m(-\omega x_{\max} \sin(\omega t + \phi_o))^2 \\
 &= \frac{1}{2}m\omega^2 x_{\max}^2 \sin^2(\omega t + \phi_o) \\
 &= \frac{1}{2}m\frac{k}{m}x_{\max}^2 \sin^2(\omega t + \phi_o) \\
 &= \frac{1}{2}kx_{\max}^2 \sin^2(\omega t + \phi_o)
 \end{aligned}$$

The potential energy due to a spring is given by (from your PH121 class or Statics/Dynamics)

$$U = \frac{1}{2}kx^2 \quad (1.21)$$

Again for our SHO we have

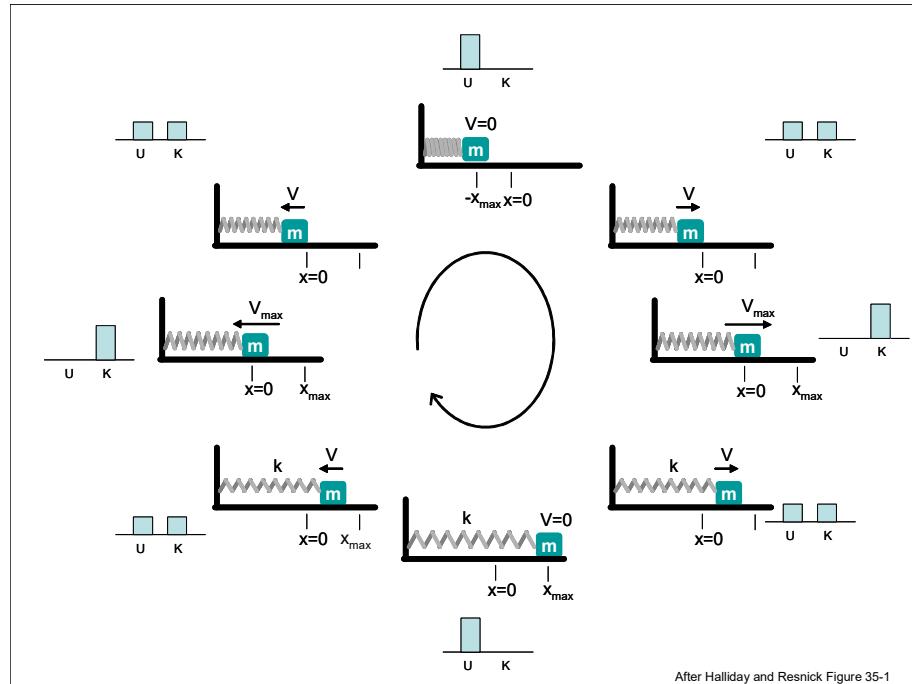
$$U = \frac{1}{2}kx_{\max}^2 \cos^2(\omega t + \phi_o) \quad (1.22)$$

The total energy is given by

$$\begin{aligned} E &= K + U \\ &= \frac{1}{2}kx_{\max}^2 \sin^2(\omega t + \phi_o) + \frac{1}{2}kx_{\max}^2 \cos^2(\omega t + \phi_o) \\ &= \frac{1}{2}kx_{\max}^2 (\sin^2(\omega t + \phi_o) + \cos^2(\omega t + \phi_o)) \\ &= \frac{1}{2}kx_{\max}^2 \end{aligned} \quad (1.23)$$

This is an astounding result! The amount of energy at any given time is equal to the amount of energy we started with. We are not changing how much energy we have. We call such a value that does not change a *constant of motion*.

Remark 1.2 *The total mechanical energy of a SHO is a constant of motion*



In the figure you can see that the kinetic and potential energies trade back and forth, but the total amount of energy does not change. Note that the kinetic and potential en-

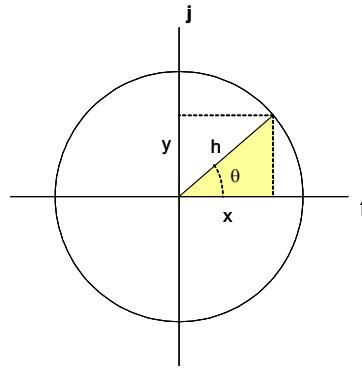
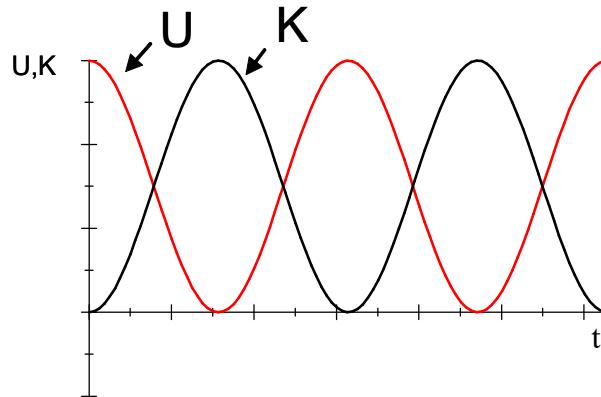


Figure 1.3.

ergy are out of phase with each other. If we plot them on the same scale (for the case $\phi_o = 0$) we have



Circular Motion and SHM

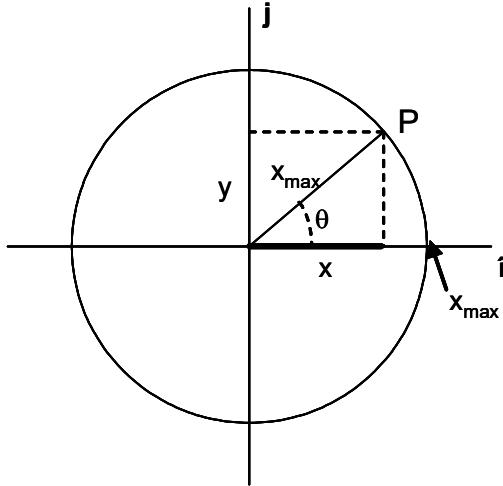
That circular motion and SHM are related should not be a surprise once we found the solutions to the equations of motion were trig functions. Recall that the trig functions are defined on a unit circle

$$\tan \theta = \frac{x}{y} \quad (1.24)$$

$$\cos \theta = \frac{x}{r} \quad (1.25)$$

$$\sin \theta = \frac{y}{r} \quad (1.26)$$

Let's relate this to our equations of motion.



Look at the projection x of the point P on the x axis. Let's follow this projection as P travels around the circle. We find it ranges from $-x_{\max}$ to x_{\max} . If we watch closely we find its velocity is zero at the extreme points and is a maximum in the middle. This projection is given as the cos of the vector from the origin to P . This model, indeed fits our SHO solution.

Now let's define a projection of P onto the y axis. Again we have SHM, but this time the projection is a sin function. Because

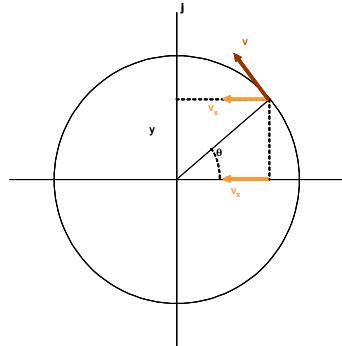
$$\cos\left(\theta - \frac{\pi}{2}\right) = \sin(\theta) \quad (1.27)$$

we can see that this is just a SHO that is 90° out of phase.

Remark 1.3 *We see that uniform circular motion can be thought of as the combination of two SHOs, with a phase difference of 90° .*

The angular velocity is given by

$$\omega = \frac{v}{r} \quad (1.28)$$



A particle traveling on the x -axis in SHM will travel from x_{\max} to $-x_{\max}$ and from $-x_{\max}$ to x_{\max} (one complete period, T) while the particle traveling with P makes one complete revolution. Thus, the angular frequency ω of the SHO and the angular velocity of the particle at P are the same. (Now we know why we used the same symbol). The magnitude of the velocity is then

$$v = \omega r = \omega x_{\max} \quad (1.29)$$

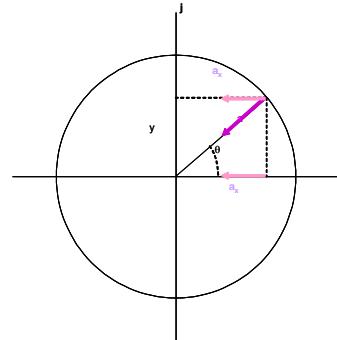
and the projection of this velocity onto the x -axis is

$$v_x = -\omega x_{\max} \sin(\omega t + \phi_o) \quad (1.30)$$

Just what we expected!

The angular acceleration of a particle at P is given by

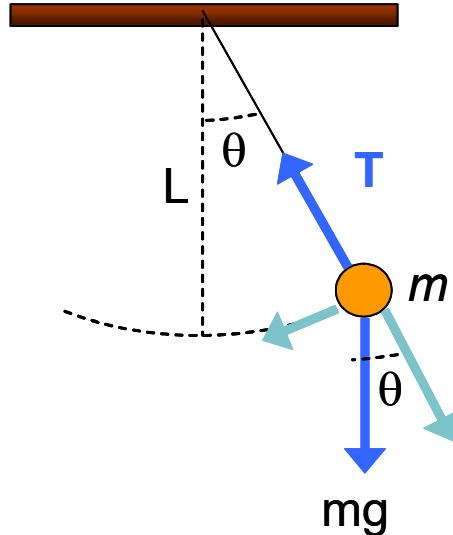
$$\frac{v^2}{r} = \frac{v^2}{x_{\max}} = \frac{\omega^2 x_{\max}^2}{x_{\max}} = \omega^2 x_{\max} \quad (1.31)$$



The direction of the acceleration is inward toward the origin. If we project this onto the x -axis we have

$$a_x = -\omega^2 x_{\max} \cos(\omega t + \phi_o) \quad (1.32)$$

The Pendulum



A simple pendulum is a mass on a string. The mass is called a “bob.”

A simple pendulum exhibits periodic motion, but not exactly simple harmonic motion.

The forces on the bob, m , are \mathbf{F}_g , \mathbf{T} the tension on the string. The tangential component of \mathbf{F}_g is always directed toward $\theta = 0$. This is a restoring force!

Let's call the path the bob takes s . The path, s , is along an arc, then from Jr. High geometry⁴, we can use the arc-length formula to describe s

$$s = L\theta \quad (10.01a)$$

and we can write an equation for the restoring force that brings the bob back to its equilibrium position as

$$\begin{aligned} F_t &= -mg \sin \theta & (1.33) \\ &= m \frac{d^2 s}{dt^2} \\ &= mL \frac{d^2 \theta}{dt^2} \end{aligned}$$

or

$$\frac{d^2 \theta}{dt^2} = -\frac{g}{L} \sin \theta$$

This is a harder differential equation to solve. But suppose we are building a grandfather

⁴ From Jr. High, but if you are like me you have forgotten it until now.

clock with our pendulum, and we won't let the pendulum swing very far. Then we can take θ as a very small angle, then

$$\sin(\theta) \approx \theta \quad (1.34)$$

In this approximation

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\theta$$

and we have a differential equation we recognize! Compare to

$$\frac{d^2x}{dt^2} = -\omega^2 x \quad (1.35)$$

if

$$\omega^2 = \frac{g}{L} \quad (1.36)$$

we have all the same solutions for s that we found for x . Since ω changed, the frequency and period will now be in terms of g and L .

$$T = \frac{2\pi}{\omega} = 2\pi \sqrt{\frac{L}{g}} \quad (1.37)$$

Remark 1.4 *the period and frequency for a pendulum with small angular displacements depend only on L and g !*

Damped Oscillations

Question 223.1.7

Question 223.1.8

Suppose we add in another force

$$\mathbf{F}_d = -b\mathbf{v} \quad (1.38)$$

This force is proportional to the velocity. This is typical of viscous fluids. So this is what we would get if we place our mass-spring system (or pendulum) in air or some other fluid. We call b the damping coefficient. Now our net force is

$$\Sigma F = -kx - bv_x = ma$$

We can write the acceleration and velocity as derivatives of the position

$$-kx - b\frac{dx}{dt} = m\frac{d^2x}{dt^2}$$

This is another differential equation. It is harder to guess its solution

$$x(t) = Ae^{-\frac{b}{2m}t} \cos(\omega t + \phi_o) \quad (1.39)$$

but now our angular frequency, ω , is more complicated

$$\omega = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2} \quad (1.40)$$

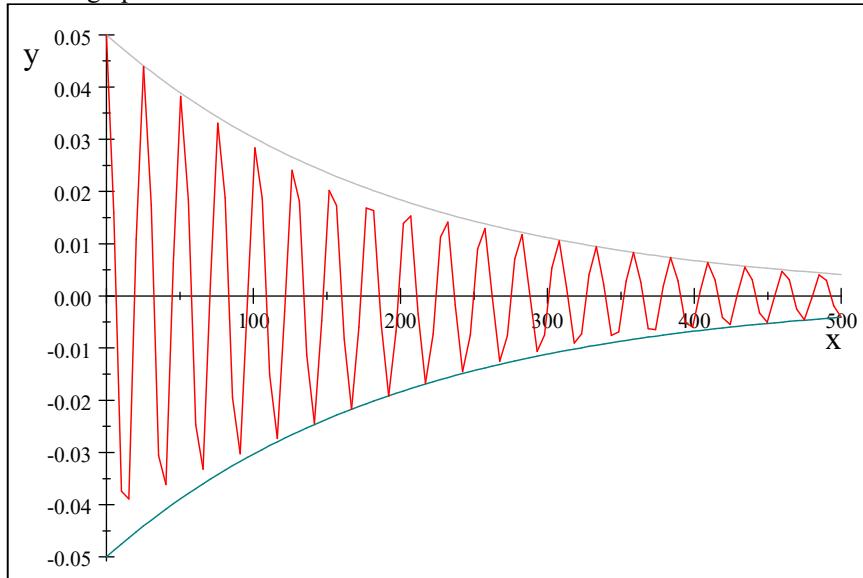
We have three cases:

- 1** 1. *the retarding force is small: ($bv_{\max} < kA$) The system oscillates, but the amplitude is smaller as time goes on. We call this “underdamped”*
2. *the retarding force is large: ($bv_{\max} > kA$) The system does not oscillate. we call this “overdamped.” We can also say that $\frac{b}{2m} > \omega_o$ (after we define ω_o below)*
3. *The system is critically damped (see below)*

For the following values,

$$\begin{aligned} A &= 5 \text{ cm} \\ b &= 0.005 \frac{\text{kg}}{\text{s}} \\ k &= .5 \frac{\text{N}}{\text{m}} \\ m &= .5 \text{ kg} \end{aligned}$$

we have a graph that looks like this



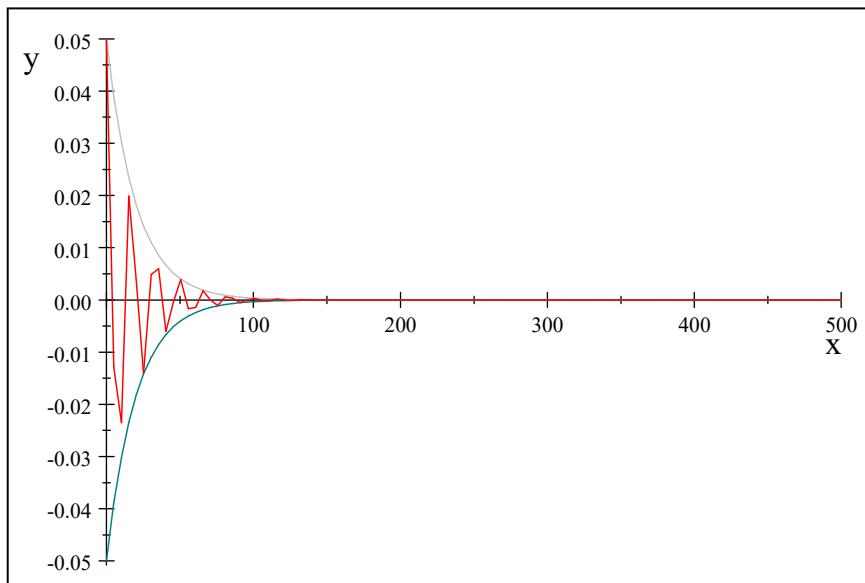
The gray lines are

$$\pm Ae^{-\frac{b}{2m}t} \quad (1.41)$$

They describe how the amplitude changes. We call this the *envelope* of the curve.

$$\begin{aligned} A &= 5 \text{ cm} \\ b &= 0.05 \frac{\text{kg}}{\text{s}} \\ k &= .5 \frac{\text{N}}{\text{m}} \\ m &= .5 \text{ kg} \end{aligned}$$

26 Chapter 1 Where We Start

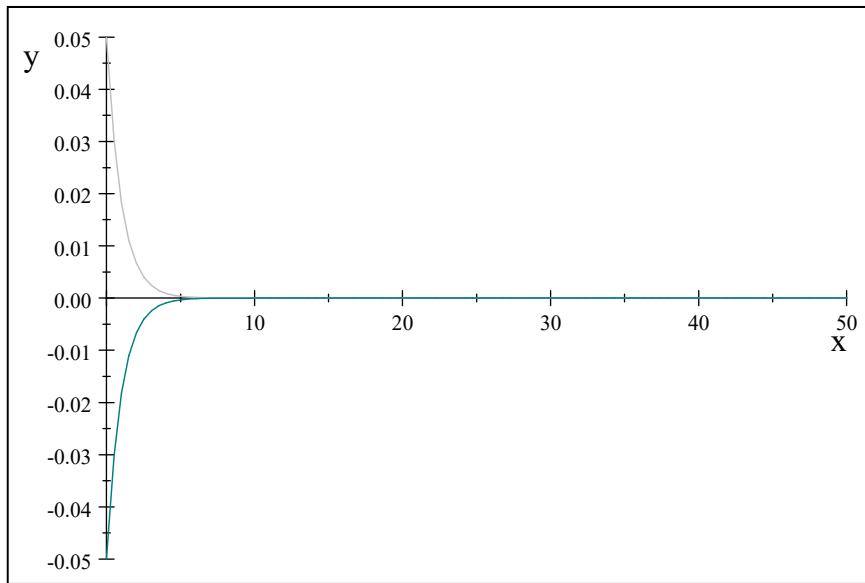


$$A = 5 \text{ cm}$$

$$b = 0.5 \frac{\text{kg}}{\text{s}}$$

$$k = .5 \frac{\text{N}}{\text{m}}$$

$$m = .5 \text{ kg}$$



What happened?

When the damping force gets bigger, the oscillation eventually stops. Only the exponential decay is observed. This happens when

$$\frac{b}{2m} = \sqrt{\frac{k}{m}} \quad (1.42)$$

then

$$\omega = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2} = 0 \quad (1.43)$$

We call this situation we call critically damped. We are just on the edge of oscillation.

We define

$$\omega_o = \sqrt{\frac{k}{m}} \quad (1.44)$$

as the *natural frequency* of the system. Then the value of b that gives us critically damped behavior is

$$b_c = 2m\omega_o \quad (1.45)$$

Remark 1.5 When $\frac{b}{2\pi} \geq \omega_o$ the solution in equation (1.39) is not valid! If you are a mechanical engineer you will find out more about this situation in your advanced mechanics classes.

Driven Oscillations and Resonance

Question 223.1.9

Question 223.1.10

We found in the last section that if we added a force like

Question 223.1.11

$$\mathbf{F}_d = -b\mathbf{v} \quad (1.46)$$

our oscillation died out. Suppose we want to keep it going? Let's apply a periodic force like

$$F(t) = F_o \sin(\omega_f t)$$

where ω_f is the angular frequency of this new driving force and where F_o is a constant.

$$\Sigma F = F_o \sin(\omega_f t) - kx - bv_x = ma$$

When this system starts out, the solutions is very messy. It is so messy that we will not give it in this class! But after a while, a steady-state is reached. In this state, the energy added by our driving force $F_o \sin(\omega_f t)$ is equal to the energy lost by the drag force, and we have

$$x(t) = A \cos(\omega_f t + \phi_o) \quad (1.47)$$

our old friend! BUT NOW

$$A = \frac{\frac{F_o}{m}}{\sqrt{\left(\omega_f^2 - \omega_o^2\right)^2 + \left(\frac{b\omega_f}{m}\right)^2}} \quad (1.48)$$

and where

$$\omega_o = \sqrt{\frac{k}{m}} \quad (1.49)$$

as before. It is more convenient to drop the f subscripts

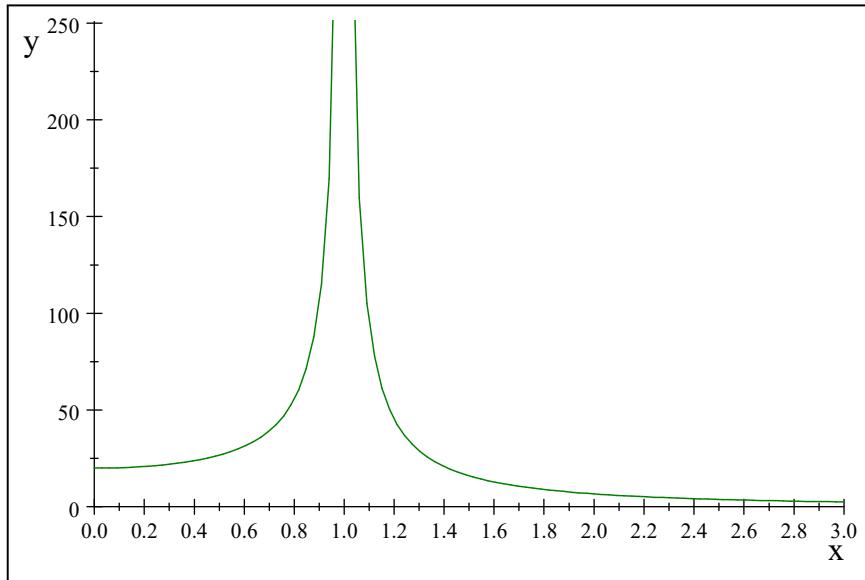
$$x(t) = A \cos(\omega t + \phi_o) \quad (1.50)$$

$$A = \frac{\frac{F_o}{m}}{\sqrt{\left(\omega^2 - \omega_o^2\right)^2 + \left(\frac{b\omega}{m}\right)^2}} \quad (1.51)$$

so now our solution looks more like our original SHM solution (except for the wild formula for A).

Lets look at A for some values of ω . I will pick some nice numbers for the other values.

$$\begin{aligned} F_o &= 2 \text{ N} \\ b &= 0.5 \frac{\text{kg}}{\text{s}} \\ k &= 0.5 \frac{\text{N}}{\text{m}} \\ m &= 0.5 \text{ kg} \end{aligned}$$



now let's calculate ω

$$\begin{aligned} \omega_o &= \sqrt{\frac{0.5 \frac{\text{N}}{\text{m}}}{0.5 \text{ kg}}} \\ &= \frac{1.0}{\text{s}} \end{aligned}$$

Notice that right at ω our solution gets very big. This is called *resonance*. To see why this happens, think of the velocity

$$\frac{dx(t)}{dt} = -\omega A \sin(\omega t + \phi_0) \quad (1.52)$$

note that our driving force is

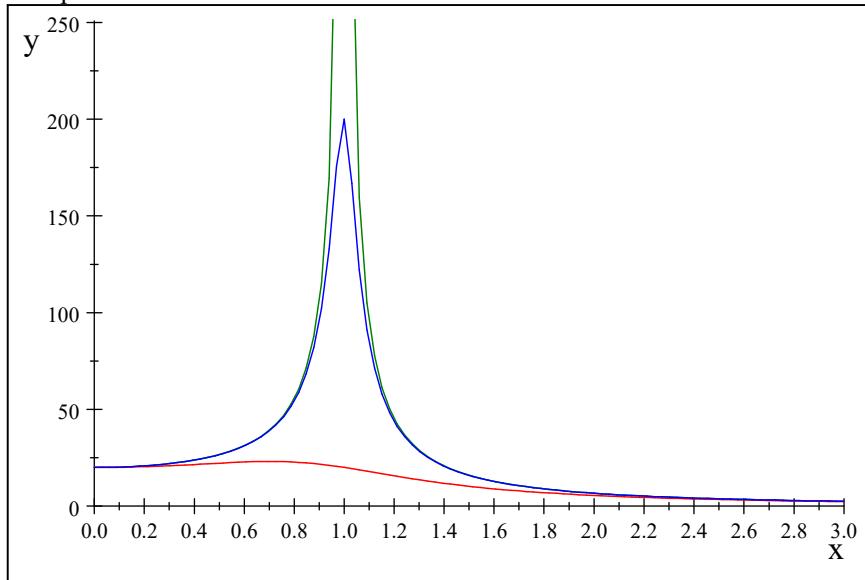
$$F(t) = F_o \sin(\omega t) \quad (1.53)$$

The rate at which work is done (power) is

$$\mathcal{P} = \frac{\mathbf{F} \cdot \Delta \mathbf{x}}{\Delta t} = \mathbf{F} \cdot \mathbf{v} \quad (1.54)$$

if F and v are in phase, the power will be at a maximum!

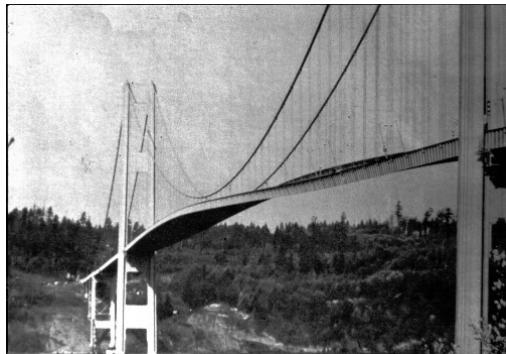
We can plot A for several values of b



Green: $b=0.005\text{kg/s}$; Blue: $b=0.05\text{kg/s}$; Red $b=0.01\text{ kg/s}$

As $b \rightarrow 0$ we see that our resonance peak gets larger. In real systems b can never be zero, but sometimes it can get small. As $b \rightarrow$ large, the resonance dies down and our A gets small.

An example of this is well known to mechanical engineers. The next picture is of the Tacoma Narrows Bridge. As a steady wind blew across the bridge it formed turbulent wind gusts.



Tacoma Narrows Bridge (Image in the Public Domain)

The wind gusts formed a periodic driving force that allowed a driving harmonic oscillation to form. Since the bridge was resonant with the gust frequency, the amplitude grew until the bridge materials broke.

2 What is a Wave?

Fundamental Concepts

1. A wave requires a disturbance, and a medium that can transfer energy
2. Waves are categorized as longitudinal or transverse (or a combination of the two).

What is a Wave?

Spring Demo Waves are organized motions in a medium.

Criteria for being a wave

Question 223.2.1

Question 223.2.2 Another way to think about waves is a transfer of energy through space without transfer of matter.

Spring Demo-
marked part

Waves require:

1. some source of disturbance
2. a medium that can be disturbed
3. some physical mechanism by which the elements of the medium can influence each other

In the limit that the string mass is negligible we represent a one-dimensional wave mathematically as a function of two variables, position and time, $y(x, t)$. There are two ways to look at waves, we call them “snapshot” and “history” (or video) views.

Longitudinal vs. transverse

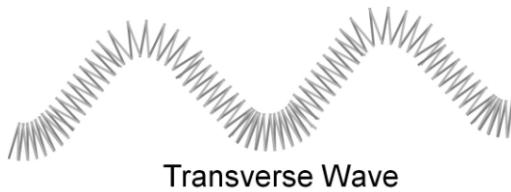
Question 223.2.3 We divide the various kinds of waves that occur into two basic types:

Question 223.2.4

Question 223.2.5

Definition 2.1 *transverse wave: a traveling wave or pulse that causes the elements of the disturbed medium to move perpendicular to the direction of propagation*

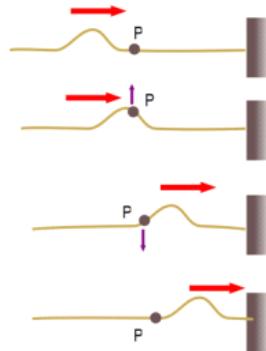
Definition 2.2 *Longitudinal wave: a traveling wave or pulse that causes the elements of the medium to move parallel to the direction of propagation*



Long Spring Demo

Examples of waves:

A pulse on a rope:

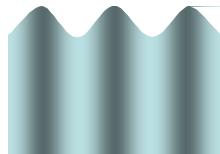


In the picture above, you see wave that has just one peak traveling to the right. We call such a wave a *pulse*. Notice how the piece of the rope marked *P* moves up and down,

but the wave is moving to the right. This pulse is a transverse wave because the parts of the medium (observe point *P*) move perpendicular to the direction the wave is moving.

An ocean wave:

Of course, some waves are a combination of these two basic types⁵. Water waves, for example, are transverse at the surface of the water, but are longitudinal throughout the water.



Earthquake waves:

Earthquakes produce both transverse and longitudinal waves. The two types of waves even travel at different speeds! *P* waves are longitudinal and travel faster, *S* waves are transverse and slower.

Wave speed

Question 223.2.6

Question 223.2.7

Question 223.2.8

We can perform an experiment with a rope or a long spring. Make a wave on the rope or spring. Then pull the rope or spring tighter and make another wave. We see that the wave on the tighter spring travels faster.

It is harder to do, but we can also experiment with two different ropes, one light and one heavy. We would find that the heavier the rope, the slower the wave. We can express this as

$$v = \sqrt{\frac{T_s}{\mu}}$$

where T_s is the tension in the rope, and μ is the linear mass density

$$\mu = \frac{m}{L}$$

where m is the mass of the rope, and L is the length.

⁵ You may have noticed that in Physics we tend to define basic types of things, and then use these basic types to define more complex objects.

The term μ might need an analogy to make it seem helpful. So suppose I have an iron bar that has a mass of 200 kg and is 2 m long. Further suppose I want to know how much mass there would be in a 20 cm section cut of the end of the rod. How would I find out?

This is not very hard, We could say that there are 200 kg spread out over 2 m, so each meter of rod has 100 kg of mass, that is, there is 100 kg/m of mass per unit length.

Then to find how much mass there is in a 0.20 m section of the rod I take

$$m = 100 \frac{\text{kg}}{\text{m}} \times 0.20 \text{ m} = 20.0 \text{ kg}$$

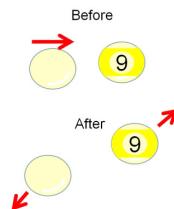
The 100 kg/m is μ . It is how much mass there is in a unit length segment of something. In this example, it is a unit length of iron bar, but for waves on string, we want the mass per unit length of string.

If you are buying stock steel bar, you might be able to buy it with a mass per unit length. If the mass per unit length is higher then the bar is more massive. The same is true with string. The larger μ , the more massive equal string segments will be.

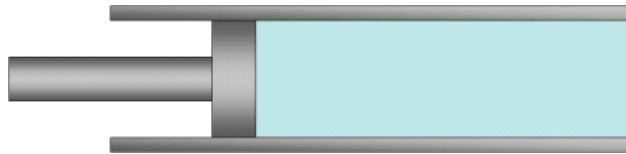
We should note that in forming this relationship, we have used our standard introductory physics assumption that the mass of the rope can be neglected. Let's consider what would happen if this were not true. Say we make a wave in a heavy cable that is suspended. The mass at the lower end of the cable pulls down on the upper part of the cable. The tension will actually change along the length of the cable, and so will the wave speed. Such a situation can't be represented by a single wave speed. But for our class, we will assume that any such changes are small enough to be ignored.

Example: Sound waves

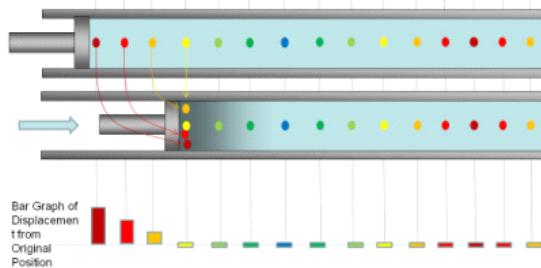
Sound is a wave. The medium is air particles. The transfer of energy is done by collision.



The wave will be a longitudinal wave. Let's see how it forms. We can take a tube with a piston in it.

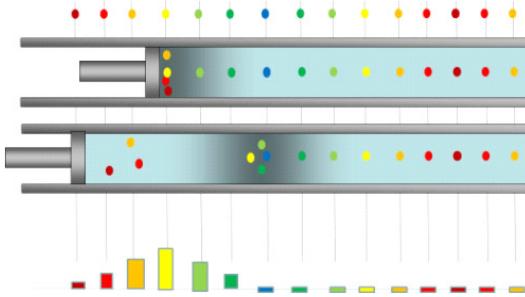


As we exert a force on the piston, the air molecules are compressed into a group. In the next figure, each dot represents a group of air molecules. In the top picture, the air molecules are not displaced. But when the piston moves, the air molecules receive energy by collision. They bunch up. We see this in the second picture.

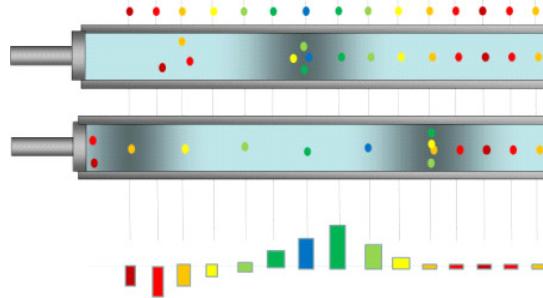


The graph below the two pictures shows how much displacement each molecule group experiences.

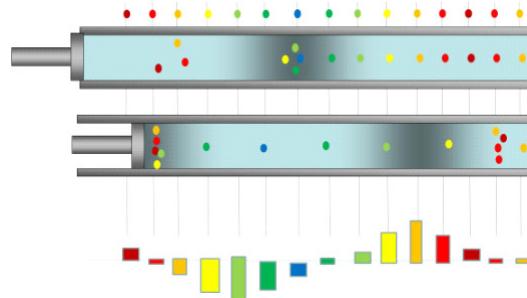
Suppose we now pull the piston back. This would allow the molecules to bounce back to the left, but the molecules that they have collided with will receive some energy and go to the right. This is shown in the next figure. Color coded dots are displayed above the before and after picture so you can see where the molecule groups started.



If we pull the piston back further, the molecules can pass their original positions.



Then we can push inward again and compress the gas.



This may seem like a senseless thing to do, but it is really what a speaker does to produce sound. In particular, a speaker is a harmonic oscillator. The simple harmonic motion of the speaker is the disturbance that makes the sound wave.



One dimensional waves

Question 223.2.9

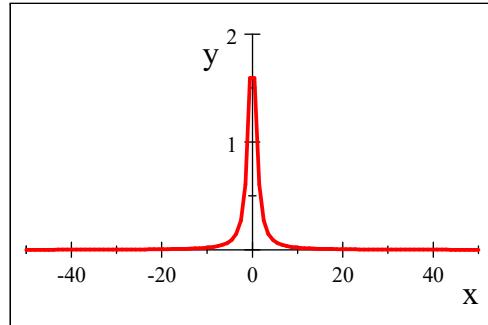
To mathematically describe a wave we will define a function of both time and position.

$$y(x, t) \quad (2.1)$$

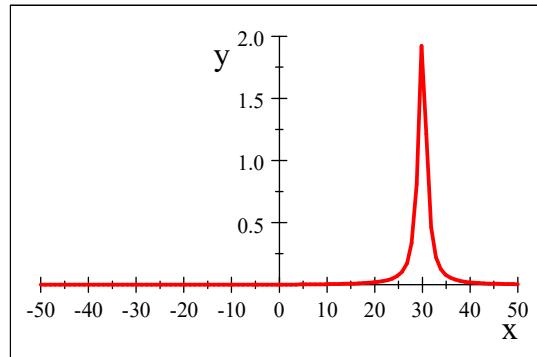
let's take a specific example⁶

$$y = \frac{2}{(x - 3.0t)^2 + 1} \quad (2.2)$$

Let's plot this for $t = 0$



what will this look like for $t = 10$?



The pulse travels along the x -axis as a function of time. We denote the speed of the pulse as v , then we can define a function

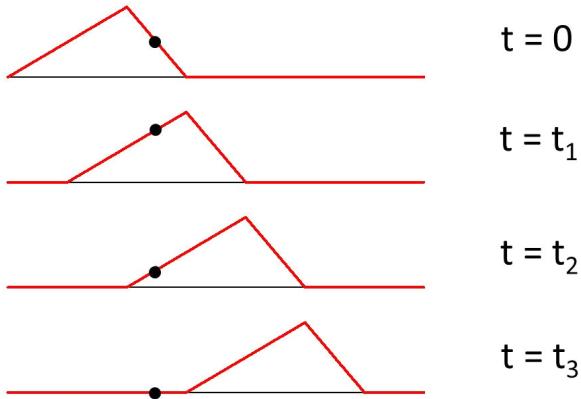
$$y(x, t) = y(x - vt, 0) \quad (2.3)$$

that describes a pulse as it travels. An element of the medium (rope, string, etc.) at position x at some time t , will have the displacement that an element had earlier at $x - vt$ when $t = 0$.

We will give $y(x - vt, 0)$ a special name, the *wave function*. It represents the y position, the transverse position in our example, of any element located at a position x at any time t .

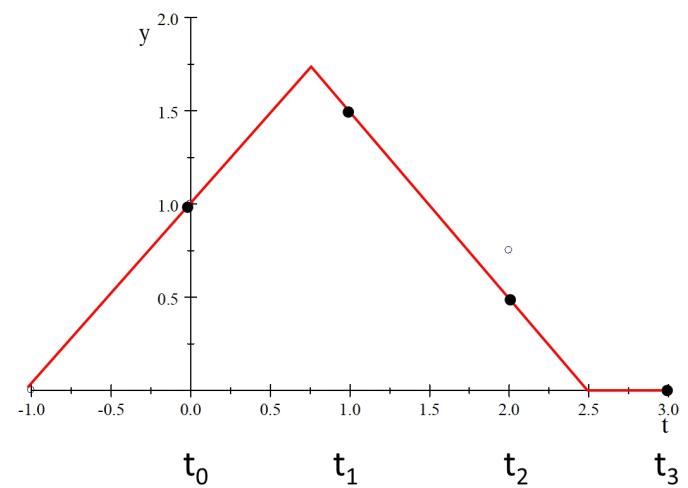
⁶ This is not an important wave function, just one I picked because it makes a nice graphic example.

Notice that wave functions depend on two variables, x , and t . It is hard to draw a wave so that this dual dependence is clear. Often we draw two different graphs of the same wave so we can see independently the position and time dependence. So far we have used one of these graphs. A graph of our wave at a specific time, t_o . This gives $y(x, t_o)$. This representation of a wave is very like a photograph of the wave taken with a digital camera. It gives a picture of the entire wave, but only for one time, the time at which the photograph was taken. Of course we could take a series of photographs, but still each would be a picture of the wave at just one time.



Question 223.2.10

The second representation is to observe the wave at just one point in the medium, but for many times. This is very like taking a video camera and using it to record the displacement of just one part of the medium for many times. You could envision marking just one part of a rope, and then using the video recorder to make a movie of the motion of that single part of the rope. We could then go frame by frame through the video, and plot the displacement of our marked part of the rope as a function of time. Such a graph is sometimes called a history graph of the wave.



3 Waves in One and More Dimensions

We studied waves in general last lecture. This time we will look at a specific wave, the sinusoidal wave. You might think this is terribly restrictive, but we will find that using sinusoidal waves, we can represent most any wave through an elegant mathematical trick, and the idea of superposition (that we will explain later).

Fundamental Concepts

1. The mathematical form of a sinusoidal wave is $y(x, t) = y_{\max} \cos(kx - \omega t + \phi_0)$
2. There are names for parts of a sinusoidal wave. We need to recognize the following terms: crest, trough, wavelength, period, frequency, angular frequency, phase constant, wave number.
3. Spatial frequency is “how often” something happens along some length.
4. The phase of a sinusoidal wave is given by $\phi = kx - \omega t + \phi_0$
5. Spherical waves have the form $y = A(r) \sin(kr - \omega t + \phi_0)$
6. Sufficiently far from the source of a wave, we can treat spherical waves like plane waves.

Sinusoidal Waves

A sinusoidal graph should be familiar from our PH121 or Dynamics experience. We can use what we know from oscillation to understand the equation for a sinusoidal wave. Remember that for simple harmonic oscillators we used the function

$$y(t) = A \cos(\omega t + \phi_0) \quad SHM \quad (3.1)$$

but this only gave us a vertical displacement at one x -position. Now our sinusoidal

function must also be a function of position along the wave.

$$y(x, t) = A \cos(kx - \omega t + \phi_0) \quad \text{waves} \quad (3.2)$$

But before we study the nature of this function, let's see what we can learn from the graph of a sinusoidal wave. We will need both of our two views, the camera snapshot and the video (history) of a point. Look at figure 3.4. This is two camera snapshots

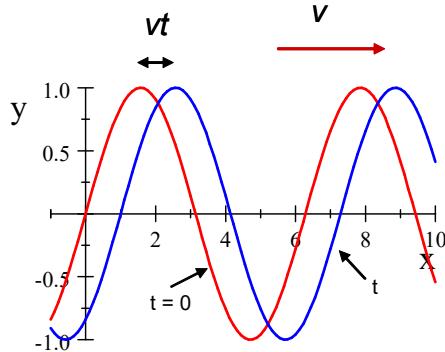


Figure 3.4.

superimposed. The red curve shows the wave (y position for each value of x) at $t = 0$. At some later time t , the wave pattern has moved to the right as shown by the blue curve. The shift is by an amount $x = vt$. This reminds us that we can write a wave function in the form

$$y(x - vt, 0)$$

Parts of a wave

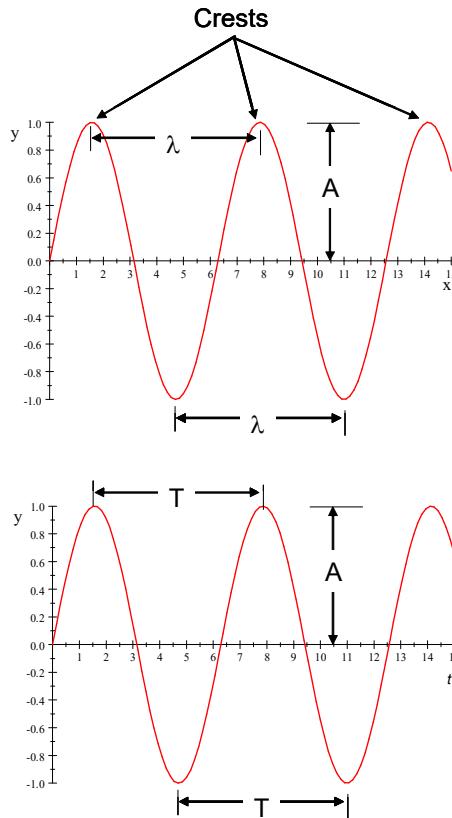
Question 123.3.1

Question 123.3.2

The peaks in the wave are called crests. For a sine wave we have a series of crests. We define the wavelength as the distance between any two nearest identical points (e.g. crests) on the wave.

Notice that this is very similar to the definition of the period, T , when we graphed SHM on a y vs. t set of axes. In fact, this similarity is even more apparent if we plot a sinusoidal wave using our two wave pictures. In the next figure, the snap-shot comes first. We can see that there will be crests. The distance between the crests is given the name *wavelength*. We give it the symbol λ . This is not the entire length of the whole wave. But it is a characteristic length of part of the wave that is easy to identify. The

next figure shows all this using our snapshot and history graphs for a sine wave.



Note that there are crests in the history graph view as well. That is because one marked part of the medium is being displaced as a function of time (think of our marked piece of the rope going up and down, or think of floating in the ocean at one point, you travel up and down as the waves go by). But now the horizontal axis is time. There will be a characteristic time between crests. That time is called the *period*. Like the wavelength is not the length of the whole wave, the period is not the time the whole wave exists. It is just the time it takes the part of the medium we are watching to go through one complete cycle. Notice that this video picture is exactly the same as a plot of the motion of a simple harmonic oscillator! For a sinusoidal wave, each part of the medium experiences simple harmonic motion.

We remember frequency from simple harmonic motion. But now we have a wave, and the wave is moving. We can extend our view of frequency by defining it as follows:

Definition 3.1 *The frequency of a periodic wave is the number of crests (or any other point of the wave) that pass a given point in a unit time interval.*

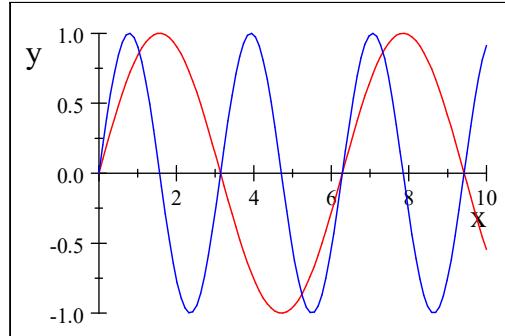


Figure 3.5.

In figure 3.5, the blue curve has twice the frequency as the red curve. Notice how it has two crests for every red crest. The maximum displacement of the wave is called the *amplitude* just as it was for simple harmonic oscillators.

Question 123.3.3

Wavenumber and wave speed

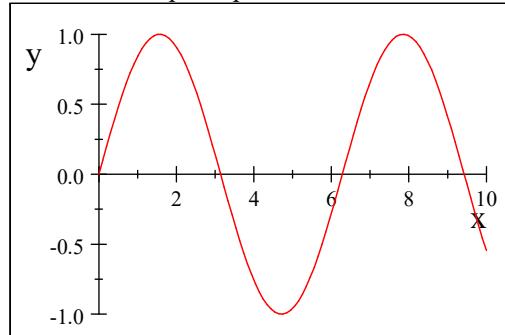
Question 123.3.4

Consider again a sinusoidal wave.

Question 123.3.5

$$y(x, t) = A \cos(kx - \omega t + \phi_0) \quad (3.3)$$

We have drawn the wave in the snapshot picture mode



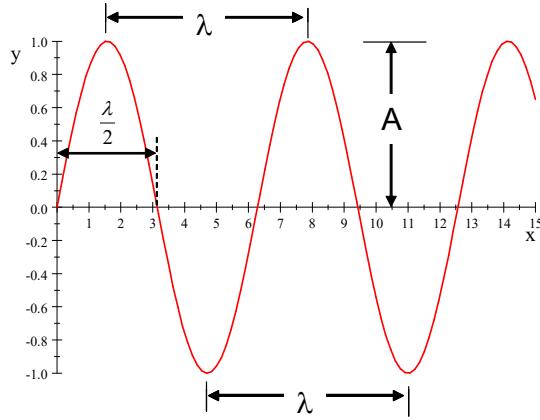
To make this graph, we set $t = 0$ and plot the resulting function

$$y(x, 0) = A \sin(kx + 0) \quad (3.4)$$

A is the amplitude. I want to investigate the meaning of the constant k . Lets find k like we did for SHM when we found ω . Consider the point $x = 0$. At this point

$$y(0, 0) = A \sin(k(0)) = 0 \quad (3.5)$$

The next time $y = 0$ is when $x = \frac{\lambda}{2}$



then

$$y\left(\frac{\lambda}{2}, 0\right) = A \sin\left(k \frac{\lambda}{2}\right) = 0 \quad (3.6)$$

From our trigonometry experience, we know that this is true when

$$k \frac{\lambda}{2} = \pi \quad (3.7)$$

solving for k gives

$$k = \frac{2\pi}{\lambda} \quad (3.8)$$

Then we now have a feeling for what k means. It is 2π over the spacing between the crests. The 2π must have units of radians attached. Then

$$y(x, 0) = A \sin\left(\frac{2\pi}{\lambda} x + 0\right) \quad (3.9)$$

We have a special name for the quantity k . It is called the *wave number*.

$$k \equiv \frac{2\pi}{\lambda} \quad (3.10)$$

Both the name and the symbol are somewhat unfortunate. Neither gives much insight into the meaning of this quantity. But from what we have done, we can understand it better. For a harmonic oscillator, we know that

$$y(t) = A \sin(\omega t)$$

where

$$\omega = 2\pi f = \frac{2\pi}{T}$$

Question 123.3.6

T is how far, in time, the crests are apart, and the inverse of this, $\frac{1}{T}$ is the frequency.

The frequency tells us how often we encounter a crest as we march along in time. So $\frac{1}{T}$ must be how many crests we have in a unit amount of time.

Now think of the relationship between the snapshot and the video representation for a

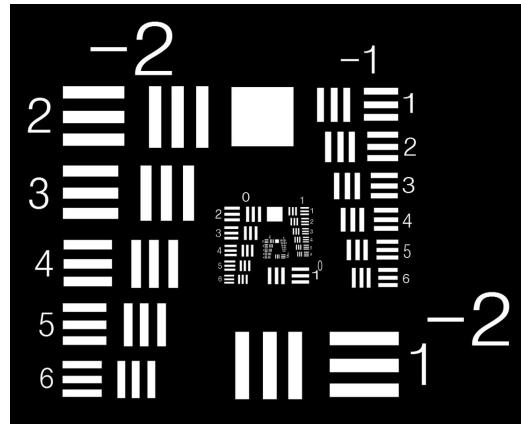
sinusoidal wave. We have a new quantity

$$k = \frac{2\pi}{\lambda}$$

where λ is how far, in distance, the crests are apart. This implies that $\frac{1}{\lambda}$ plays the same role in the snap shot graph that f plays in the video graph. It must tell us how many crests we have, but this time it is how many crests in a unit of distance. We found above that k told us something about how often the zeros (well, every other zero) will occur. But the crests must occur at the same rate. So k tells us how often we encounter a crest in our snapshot graph.

The frequency is how often we encounter a crest in the video graph, $\frac{1}{\lambda}$ is how often we encounter a crest in the snap shot graph. Thus $\frac{1}{\lambda}$ is playing the same role for a snap shot graph as frequency plays for a history graph. We could call $\frac{1}{\lambda}$ a *spatial frequency*. It is how often we encounter a crest as we march along in position, or how many crests we have in a unit amount of distance. And λ could be called a *spatial period*. Both $1/T$ and $1/\lambda$ answer the question “how often something happens in a unit of something” but one asks the question in time and the other in position along the wave.

My mental image for this is the set of groves on the edge of a highway. There is a distance between them, like a wavelength, and how often I encounter one as I move a distance along the road is the spatial frequency. If you are a farmer, you may think of plowed fields, with a distance between furrows as a wavelength, and how closely spaced the furrows are as a measure of spatial frequency. We use this concept in optics to test how well an optical system resolves details in a photograph. The next figure is a test image. A good camera will resolve all spatial frequencies equally well. Notice the test image has sets of bars with different spatial frequencies. By forming an image of this pattern, you can see which spacial frequencies are faithfully represented by the optical system.



Resolution test target based on the USAF 1951 Resolution Test Pattern (not drawn to exact specifications).

In class you will see that our projector does not represent all spatial frequencies equally well! You can also see this now in the copy you are reading. If you are reading on-line or an electronic copy, your screen resolution will limit the representation of some spatial frequencies. Look for the smallest set of three bars where you can still tell for sure that there are three bars. A printed version that has been printed on a laser printer will usually allow you to see even smaller sets of three bars clearly.

Let's place k in the full equation for the sine wave for any time, t .

$$y(t) = A \cos(kx - \omega t + \phi_o) \quad (3.11)$$

We would like this to look like our wave function equation

$$y(x, t) = y(x - vt, 0)$$

With a little algebra we can do this

$$\begin{aligned} y(t) &= A \cos(kx - \omega t + \phi_o) \\ &= A \cos\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{T}t + \phi_o\right) \\ &= A \cos\left(\frac{2\pi}{\lambda}\left(x - \frac{\lambda}{T}t\right) + \phi_o\right) \end{aligned}$$

This is in the form of a wave function so long as

$$v = \frac{\lambda}{T} \quad (3.12)$$

then

$$y(x, t) = A \sin\left(\frac{2\pi}{\lambda}(x - vt) + \phi_o\right) \quad (3.13)$$

We can see that the wave travels one wavelength in one period. The simple relationship

$$v = \frac{\lambda}{T} \quad (3.14)$$

is of tremendous importance.

Wave speed forms

We found

$$v = \frac{\lambda}{T} \quad (3.15)$$

but it is easy to see that

$$v = \frac{2\pi\lambda}{2\pi T} = \frac{\omega}{k} \quad (3.16)$$

and

$$v = \lambda f \quad (3.17)$$

This last formula is, perhaps, the most common form encountered in our study of light.

Phase

You may be wondering about the phase constant we learned about in our study of SHM. We have ignored it up to now. But of course we can shift our sine just like we did for our plots of position vs. time for oscillation. Only now with a wave we have two graphs, a history and snapshot graphs, so we could shift along the x in a snapshot graph or along the t axes in a history graph. So the sine wave has the form.

$$y(x, t) = A \sin(kx - \omega t + \phi_0) \quad (3.18)$$

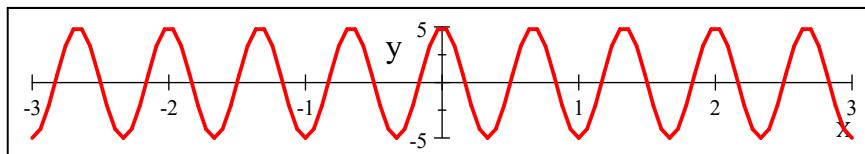
where ϕ_0 will need to be determined by initial conditions just like in SHM problems and those initial conditions will include initial positions as well as initial times.

Let's consider that we have two views of a wave, the snapshot and history view. Each of these looks like sinusoids for a sinusoidal wave. Let's consider a specific wave,

$$y(x, t) = 5 \sin\left(3\pi x - \frac{\pi}{5}t + \frac{\pi}{2}\right)$$

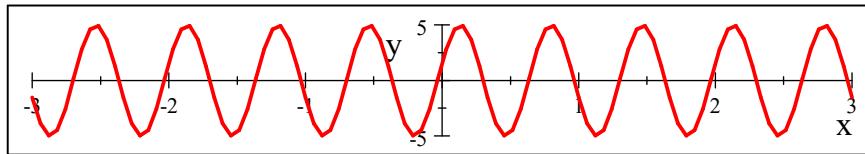
And let's look at a snapshot graph at $t = 0$

$$y(x, 0) = 5 \sin\left(3\pi x - \frac{\pi}{5}(0) + \frac{\pi}{2}\right)$$



and another at $t = 2$ s

$$y(x, 2\text{ s}) = 5 \sin\left(3\pi x - \frac{\pi}{5}(2\text{ s}) + \frac{\pi}{2}\right)$$



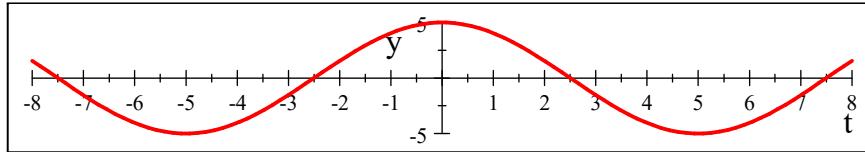
Comparing the two, we could view the latter as having a different phase constant

$$\phi_{total} = \omega\Delta t + \phi_o = -\frac{\pi}{5}(2\text{ s}) + \frac{\pi}{2}$$

that is, within the snapshot view, the time dependent part of the argument of the sine acts like an additional phase constant.

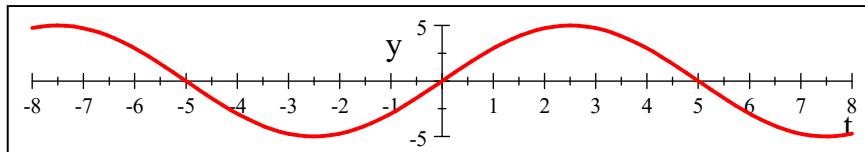
Likewise, in the history view, we can plot our wave at $x = 0$

$$y(0, t) = 5 \sin\left(3\pi(0) - \frac{\pi}{5}t + \frac{\pi}{2}\right)$$



and at $x = 1.5\text{ m}$

$$y(1.5\text{ m}, t) = 5 \sin\left(3\pi(1.5\text{ m}) - \frac{\pi}{5}t + \frac{\pi}{2}\right)$$



Within the history view, the kx part of the argument acts like a phase constant.

$$\phi_{total} = k\Delta x + \phi_o = 3\pi(1.5\text{ m}) + \frac{\pi}{2}$$

Of course neither kx nor ωt are constant, But within individual views of the wave we have set them as constant to form our snapshot and history representations. We can see that any part of the argument of the sine, $kx - \omega t + \phi_o$ could contribute to a phase shift, depending on the view we are taking.

Because of this, it is customary to call the entire argument of the sine function, $\phi = kx - \omega t + \phi_o$ the *phase of the wave*. Where ϕ_o is the phase constant, ϕ is the phase. Of course then, ϕ must be a function of x and t , so we have a different value for $\phi(x, t)$ for every point on the wave for every time.

Sinusoidal waves on strings

Take a jump rope, and shake one end up and down while your partner keeps his or her end stationary. You can make a sine wave in the rope. You can do a better job by attaching a wave generator to the end.

Really, as long as the wave forms are identical and periodic, the relationships

$$f = \frac{1}{T} \quad (3.19)$$

and

$$v = f\lambda \quad (3.20)$$

will hold. But we will make our device vibrate with simple harmonic motion.

Let's call an element of the rope Δx . Here the “ Δ ” is being used to mean “a small amount of.” We are taking a small amount of the rope and calling it's length Δx .

Each element of the rope (Δx_i) will also oscillate with SHM (think of a driven SHO). Note that the elements of the rope oscillate in the y direction, but the wave travels in x . This is a transverse wave.

Let's describe the motion of an element of the string at point P .

At $t = 0$,

$$y = A \sin(kx - \omega t) \quad (3.21)$$

(where I have chosen $\phi_0 = 0$ for this example). The element does not move in the x direction. So we define the *transverse speed*, v_y , and the *transverse acceleration*, a_y , as the velocity and acceleration of the element of rope in the y direction. These are not the velocity and acceleration of the wave, just the velocity and acceleration of the element Δx at a point P .

Because we are doing this at one specific x location we need partial derivatives to find the velocity

$$v_y = \left. \frac{dy}{dt} \right|_{x=\text{constant}} = \frac{\partial y}{\partial t} \quad (3.22)$$

That is, we take the derivative of y with respect to t , but we pretend that x is not a variable because we just want one x position. Then

$$v_y = \frac{\partial y}{\partial t} = -\omega A \cos(kx - \omega t) \quad (3.23)$$

and

$$a_y = \frac{\partial v_y}{\partial t} = -\omega^2 A \sin(kx - \omega t) \quad (3.24)$$

These solutions should look very familiar! We expect them to be the same as a harmonic oscillator except that we now have to specify which oscillator—which part of the rope—we are looking at. That is what the kx part is doing.

The speed of Waves on Strings

Only use if there are questions on this.

Let's work a problem together. Let's find an expression for the speed of the wave as it travels along a string.

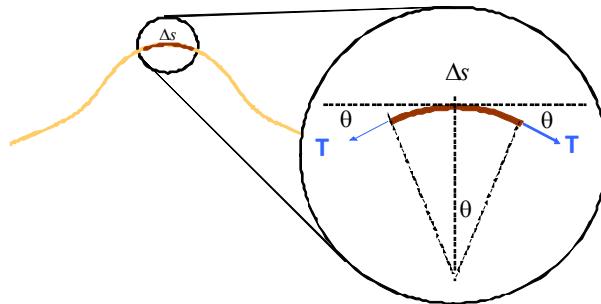


Figure 3.6.

We will use Newton's second law

$$\Sigma \vec{F} = \vec{F}_{net} = m \vec{a}$$

to do this, so we need a sum of the forces. What are the forces acting on an element of string?

- Tension on the right hand side (RHS) of the element from the rest of the string on the right, T_r
- Tension on the left hand side (LHS) of the element from the rest of the string on the left, T_l
- The force due to gravity on our element of string, F_g

Lets assume that the element of string, Δs , at the crest is approximately an arc of a circle with radius R .

There is a force pulling left on the left end of the element that is tangent to the arc, there is a force pulling right at the right end of the element which is also tangent to the arc. The horizontal components of the forces cancel ($T \cos \theta$). The vertical component, ($T \sin (\theta)$) is directed toward the center of the arc. Then these forces must be a mass times an acceleration and because they are center seeking we can call these

accelerations centripetal accelerations

$$a = \frac{v^2}{R} \quad (3.25)$$

If the rope is not moving in the x direction, then

$$\Sigma F_x = 0 = -T_l \cos \theta + T_r \cos \theta$$

$$T_l = T_r$$

Then, the radial force F_r will have matching components from each side of the element that together are $2T \sin(\theta)$. Since the element is small,

$$\Sigma F_r = 2T \sin(\theta) \approx 2T\theta \quad (3.26)$$

The element has a mass m .

$$m = \mu \Delta s \quad (3.27)$$

where μ is the mas per unit length. Using the arc length formula

$$\Delta s = R(2\theta) \quad (3.28)$$

so

$$m = \mu \Delta s = 2\mu R\theta \quad (3.29)$$

and finally we use the formula for the radial acceleration

$$F_r = ma = (2\mu R\theta) \frac{v^2}{R} \quad (3.30)$$

Combining these two expressions for F_r

$$2T\theta = (2\mu R\theta) \frac{v^2}{R} \quad (3.31)$$

$$T = (\mu R) \frac{v^2}{R} \quad (3.32)$$

$$\frac{T}{\mu} = v^2 \quad (3.33)$$

and we find that

$$v = \sqrt{\frac{T}{\mu}} \quad (3.34)$$

Note that we made many assumptions along the way. Despite this, the approximation is quite good.

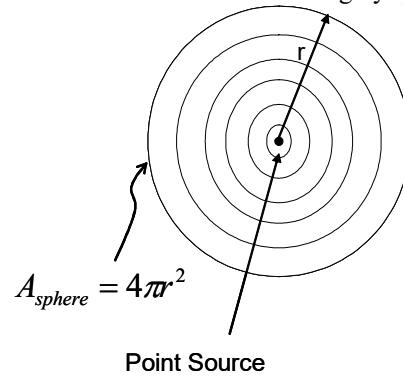
Waves in two and three dimensions

So far we have written expressions for waves, but our experience tells us that waves don't usually come as one dimensional phenomena. In the next figure, we see the disturbance (a drop) creating a water wave.



Picture of a water drop (Jon Paul Johnson, used by permission)

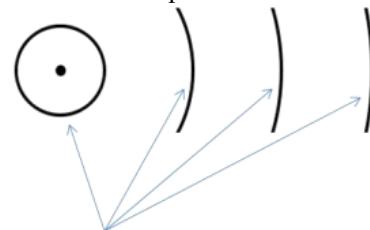
The wave is clearly not one dimensional. It appears nearly circular. In fact, it is closer to hemispherical, and this limit is only true because the disturbance is at the air-water boundary. Most waves in a uniform medium will be roughly spherical.



As such a wave travels away from the source, the energy traveling gets more spread out. This causes the amplitude to decrease. Think of a sound wave, it gets quieter the farther you are from the source. We change our equation to account for this by making the amplitude a function of the distance, r , from the source

$$y = A(r) \sin(\vec{k} \cdot \vec{r} - \omega t + \phi_0) \quad (3.35)$$

Of course, if we look at a very large wave, but we only look at part of the wave, we see that our part looks flatter as the wave expands.



Portion of a Spherical Wave: Wave becomes more flat as it expands

Very far from the source, our wave is flat enough that we can ignore the curvature across its wave fronts. We call such a wave a *plane wave*. There are no true plane waves in nature, but this idealization makes our mathematical solutions simpler and many waves come close to this approximation. We will usually stick with the plane wave approximation in this class.

Basic Equations

Our general wave equation is

$$y(x, t) = A \cos(kx - \omega t + \phi_0) \quad (3.36)$$

or

$$y(x, t) = A \sin(kx - \omega t + \phi_0) \quad (3.37)$$

In general we found that the wave speed could be written as

$$v = \frac{\omega}{k} \quad (3.38)$$

$$v = \lambda f \quad (3.39)$$

For waves on strings we also found

$$v = \sqrt{\frac{T}{\mu}} \quad (3.40)$$

We found that all of

$$\phi = kx - \omega t + \phi_0 \quad (3.41)$$

could be called the total phase of the wave, or to be brief, we could call it just the phase of the wave.

The *transverse* position, speed, and acceleration of a part of the medium are given by

$$y = A \sin(kx - \omega t) \quad (3.42)$$

$$v_y = -\omega A \cos(kx - \omega t) \quad (3.43)$$

$$a_y = -\omega^2 A \sin(kx - \omega t) \quad (3.44)$$

We found that waves in three dimensions have a more complicated amplitude

$$y = A(r) \sin(\vec{k} \cdot \vec{r} - \omega t + \phi_o) \quad (3.45)$$

4 Light, Sound, Power

Reading Assignment 20.5,20.6

Fundamental Concepts

- Sound waves are formed when a disturbance causes a chain-reaction of collisions in the molecules of the air or other substance.
- Power is an amount of energy expended in an amount of time
- Intensity is an amount of power spread over an area
- The human auditory system is not a linear , but rather a logarithmic detector with perceived sound level given by $\beta = 10 \log_{10} \left(\frac{I}{I_o} \right)$

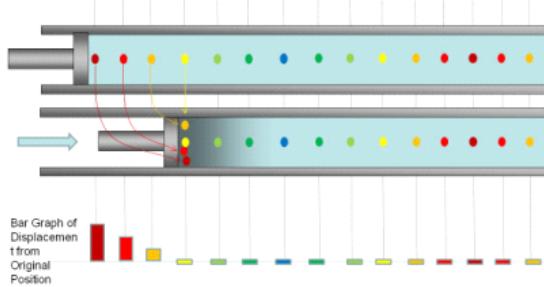
Waves in matter-Sound

We have said that sound is a longitudinal wave with a medium of air. Really any solid, liquid, or gas will work as a medium for sound. For our study, we will take sound to be a longitudinal wave and treat liquids and gasses. Solids have additional forces involved due to the tight bonding of the atoms, and therefore are more complicated. Technically in a solid sound can be a transverse wave as well a longitudinal wave, but we usually call transverse waves of this nature *shear waves*.

Question 223.4.1

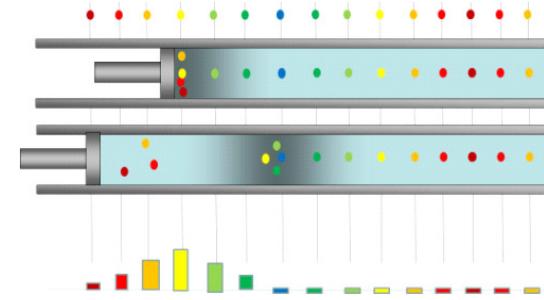
Periodic Sound Waves

Let's go back to making sounds. Suppose we push our piston as we did before.



When we push in the piston, it creates a region of higher pressure next to it.

When we pull back the piston the fluid expands to fill the void.



We create a rarefaction next to the piston.

Suppose we drive the piston sinusoidally. Can we describe the motion of the particles and of the wave?

Definition 4.1 *Compression: A local region of higher pressure in a fluid*

Definition 4.2 *Rarefaction: A local region of lower pressure in a fluid*

We can identify the distance between two compressions as λ .

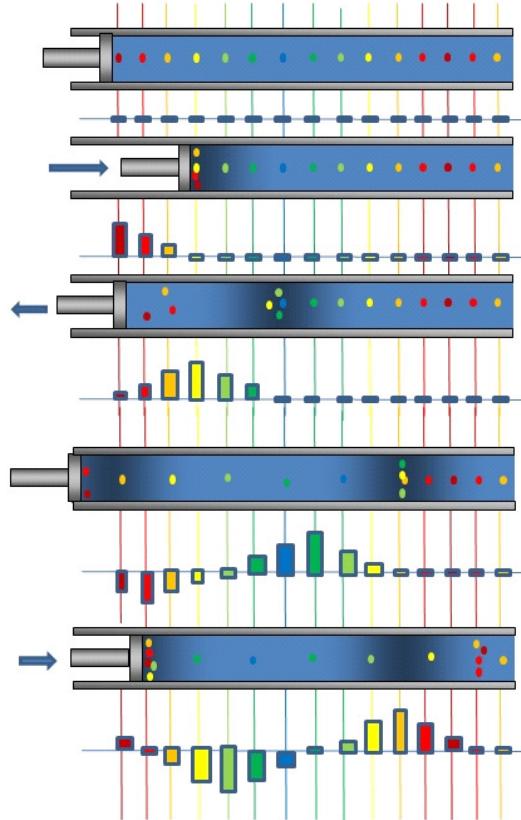
We define $s(x, t)$ like we defined a wave function, $y(x, t)$ as the displacement of a particle of fluid relative to its equilibrium position.

$$s(x, t) = s_{\max} \cos(kx - \omega t) \quad (4.1)$$

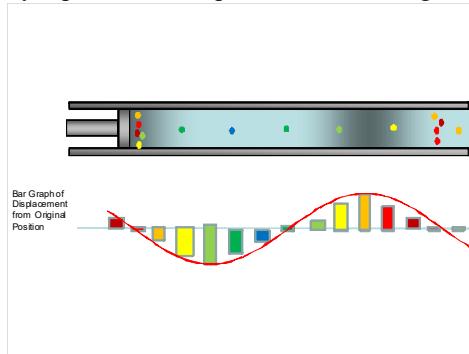
but what is s_{\max} ?

We remember that s_{\max} is the maximum displacement of a particle of fluid from its equilibrium position. We plotted this using a bar graph to show displacement from the

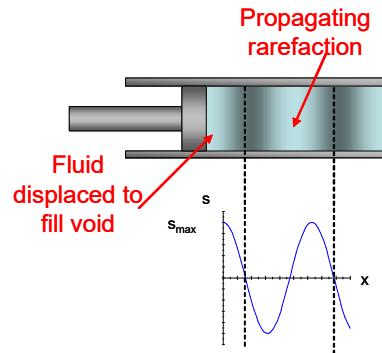
equilibrium position for our molecules. As we push the piston in and out we will get something like this.



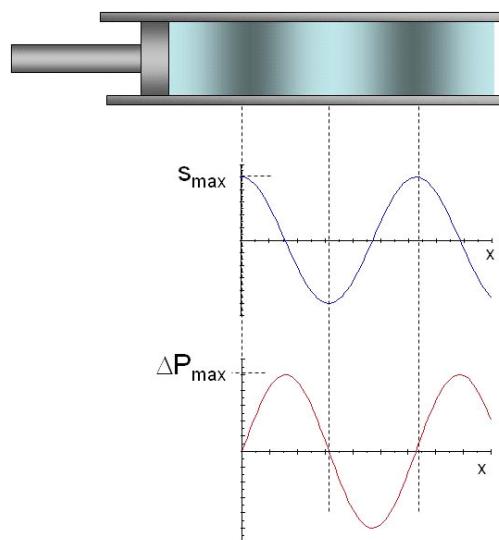
We found before that we get something that looks like a sine wave, but remember what the bars represent. They represent the displacement from original position.



We don't usually draw bar graphs to represent sound waves, we usually just draw the sine wave.



When the air molecules bunch up to form a compression, the pressure will be higher. When the air molecules spread out to form a rarefaction, the pressure will be lower than normal. The variation of the gas pressure ΔP measured from its equilibrium is also periodic



which is why we often refer to a sound wave as a pressure wave. Think of when the wave gets to your ear. The wave consists of a group of particles all headed for your ear drum. When they hit, they exert a force. Pressure is a force spread over an area,

$$P = \frac{F}{A}$$

so in a sense, we hear changes in air pressure!

Speed of Sound Waves

Question 223.4.2

The speed of sound in air is around 340 m/s. The speed changes when we change media, and even when we are in the same media but the temperature changes. For sound in air, a good approximation near standard pressure and temperature is

$$v = v_o \sqrt{1 + \frac{T_c}{T_o}} \quad (4.2)$$

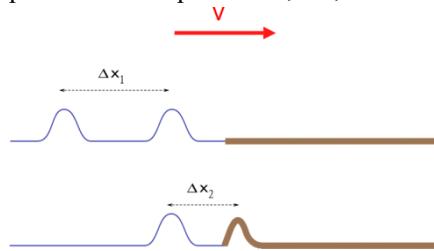
where $v_o = 331 \frac{\text{m}}{\text{s}}$ and $T_o = 273 \text{ K}$ (0°C).⁷

Why temperature? The density and pressure of air change with temperature. The air molecules gain kinetic energy and tend to move farther apart from each other when they are warm. This changes the time it takes to transfer energy.

Boundaries

Question 223.4.3

Suppose two pulses travel in the same medium, say, on a rope, and they approach a different rope with a different linear mass density. If the new rope is heavier, we expect the wave speed to slow down. So as one pulse reaches the boundary, it will go slower. This allows the second pulse to catch-up before it, too, slows down at the boundary.



Now suppose a sinusoidal wave approaches the boundary. We can envision the crests like pulses, and we expect the first crest to slow down when it reaches the boundary, letting the other crests catch up. Once the wave passes the boundary, the crests will be closer together. The wavelength changes as we move to the slower medium.

But does the frequency change? We know that

$$v = \lambda f$$

so

$$f = \frac{v}{\lambda}$$

⁷ $v = v_o \sqrt{1 + \frac{T_c}{T_o}} = v_o \sqrt{\frac{T_o}{T_o} + \frac{T_c}{T_o}} = v_o \sqrt{\frac{T_o + T_c}{T_o}} = v_o \sqrt{\frac{T_K}{T_o}}$

both the speed and the wavelength have changed, but did they change proportionately so f is constant? This must be so. Think that the change in wavelength is due to the relative speed of the wave in the two media. If Δv is small, the change in λ will be small because the crests are not delayed too long. If Δv is large, the crests are delayed by a large amount and so the change in λ is large. We won't derive the fact that f is constant, but we can see that it is very believable that it is true.

This is true for all waves, even light. When a wave crosses a boundary from a fast to a slow or a slow to a fast medium, λ will change and f will remain constant.

Question 223.4.4

Let's find an expression for the new wavelength. The frequency of the light must be the same.

$$f_i = f_f$$

and we know that in general

$$f = \frac{v}{\lambda}$$

so

$$\frac{v_i}{\lambda_i} = \frac{v_f}{\lambda_f}$$

so

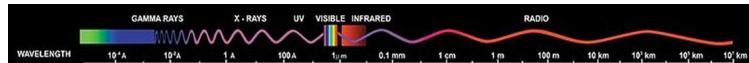
$$\lambda_f = \frac{v_f}{v_i} \lambda_i \quad (4.3)$$

and we can see that if v_f is slower than v_i our wavelength does get shorter.

Waves in fields-Light

Sound is a wave in matter, but what is light? It will really take the rest of the course (and then some) to answer this question. But we know that light can travel through a vacuum. Therefore, light can't be a wave in some type of matter. We will find later that there exists something called an electromagnetic field created by charged particles. It turns out that light seems to be a wave in this electromagnetic field. It will take us a while to fully understand this concept, but don't worry. Physicists knew that light was a wave for almost 80 years before the electric field was shown to be the medium. We can do a lot just knowing light behaves like a wave.

If light is a wave, the light we see is just one small part of a whole class of waves that are possible in this electromagnetic field medium. Radio waves, and microwaves, and x-rays are all just different types of electromagnetic waves. The next figure shows where all of these electromagnetic waves fit ordered by wavelength (and frequency).



Electromagnetic Spectrum (Public Domain image courtesy NASA)

Speed of Electromagnetic waves

There is something very unique about this electromagnetic field medium. The waves in this medium travel at a constant speed-no matter what frame of reference we are in. This fact leads to the formation of the Special Theory of Relativity and the famous equation

$$E = mc^2$$

where c is this speed of light

$$c = 299792458 \frac{\text{m}}{\text{s}}$$

Light does slow down when it enters a material medium, like glass, or even air. The actual speed that light travels does not change. What happens is that light is absorbed by the electrons in the atoms of the material substance. The electron temporarily takes up all the energy from a bit of the light wave—but only temporarily. It eventually has to give up the energy and the light wave reforms. But it has lost some time in the process, so its average speed is less. How much less depends on how long the electrons in the atoms can hang-on to the light. Each substance is different.

We can devise a way to express how much slower light will appear to go in a substance using the ratio

$$\frac{c}{v}$$

the ratio of the speed of light, c , to the average speed in the substance, v . This ratio is so useful that we give it a name, the *index of refraction*.

$$n = \frac{c}{v}$$

Power and Intensity

We know that energy is being transferred by the wave, whether it is a light or sound wave. We should wonder, how fast is energy transferring? This can mean the difference between sunlight on a warm summer day and being burned by a laser beam. We will start by considering the rate of energy transfer, *power*.

64 Chapter 4 Light, Sound, Power

Question 222.4.5

Power

Question 223.4.6

The concept of power should be familiar to us from PH121 or Statics and Dynamics.

We can find the power by

$$\mathcal{P} = \frac{\Delta E}{\Delta t}$$

where ΔE is the power transferred and Δt is the time it takes to make the transfer.

Intensity

Tuning Fork Demo

Question 223.4.7

We now define something new.

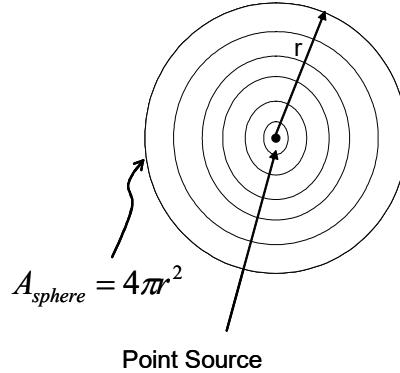
QQuestion 223.4.8

$$\mathcal{I} \equiv \frac{\mathcal{P}}{A} \quad (4.4)$$

Question 223.4.9

that is, the power divided by the area. But what does it mean?

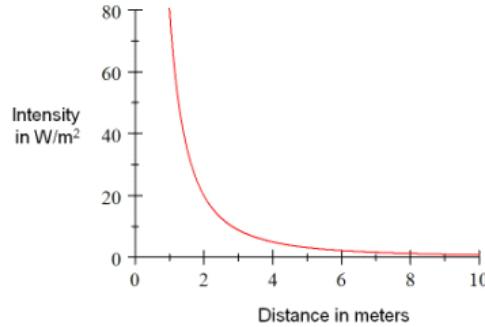
Consider a point source.



it sends out waves in all directions. The wave crests will define a sphere around the points source (the figure shows a cross section but remember it is a wave from a point source, so we are really drawing concentric spheres like balloons inside of balloons.). Then form our point source

$$\mathcal{I} = \frac{\mathcal{P}}{4\pi r^2} \quad (4.5)$$

As the wave travels, its power per unit area decreases with the square of the distance (think gravity) because the area is getting larger.



This quantity that tells us how spread our power has become is called the *intensity* of the wave.

Suppose we cup our hand to our ear. We can now hear fainter sounds. But what are we doing that makes the difference?

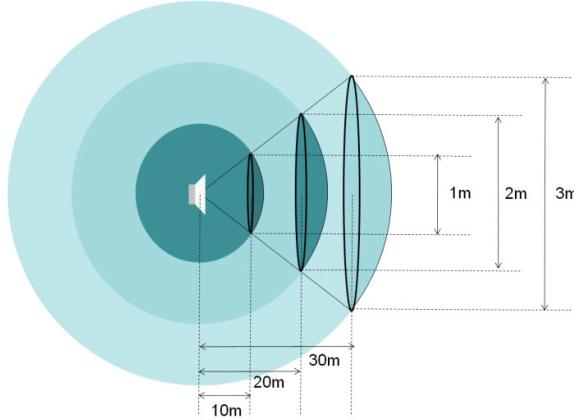
We are increasing the area of our ear. Our ears work by transferring the energy of the sound wave to a electro-chemical-mechanical device that creates a nerve signal. The more energy, the stronger the signal. If we are a distance r away from the source of the sound then the intensity is

$$\mathcal{I} = \frac{\mathcal{P}_{source}}{A_{wave}}$$

But we are collecting the sound wave with another area, the area of our hand. The power received is

$$\begin{aligned}\mathcal{P}_{received} &= \mathcal{I} A_{hand} \\ &= \frac{A_{hand}}{A_{wave}} \mathcal{P}_{source}\end{aligned}$$

and we can see that, indeed, the larger the hand, the more power, and therefore more energy we collect. This is the idea behind a dish antenna for communications and the idea behind the acoustic dish microphones we see at sporting events. In next figure, we can see that it would take an increasingly larger dish to maintain the same power gathering capability as we get farther from the source.



Sound Levels in Decibels

Our Design Engineer made an interesting choice in building us. We need to hear very faint sounds, and very loud sounds too. In order to make us able to hear the soft sounds without causing extreme discomfort when we hear the loud, He gave up linearity.

Question 223.4.10

Question 223.4.11

Question 223.4.12

Sound Meter Demo

That is, we don't hear twice the sound intensity as twice as loud. The mathematical expression that matches our perception of loudness to the intensity is

$$\beta = 10 \log_{10} \left(\frac{I}{I_o} \right) \quad (4.6)$$

where the quantity I_o is a reference intensity. We are comparing the intensity of our sound with some reference intensity, I_o , to see how much louder our sound seems to be.

We call β the *sound level*. I_o we choose to be the *threshold of hearing*, the intensity that is just barely audible. Measured this way, we say that intensity is in units of decibels (dB). The decibel, is an engineer's friend (and useful for physicists too!) because it can describe a non-linear response in a linear way that is easy to match to our human experience.

Suppose we double the intensity by a factor of 2.

$$\begin{aligned} \beta &= 10 \log_{10} \left(\frac{2I_o}{I_o} \right) \\ &= 10 \log_{10} 2 \\ &= 3.0103 \text{ dB} \end{aligned}$$

The sound intensity level is not twice as large, but only 3dB larger. It is a tiny increase.

Question 223.4.13

This is what we hear. A good rule to remember is that $3dB$ corresponds to a doubling of the intensity.

Question 223.4.14

The tables that follow give some common sounds in units of dB and W/m^2 . Just for reference, I have measured a Guns n Roses concert at 120 dB outside the stadium.

Question 223.4.15

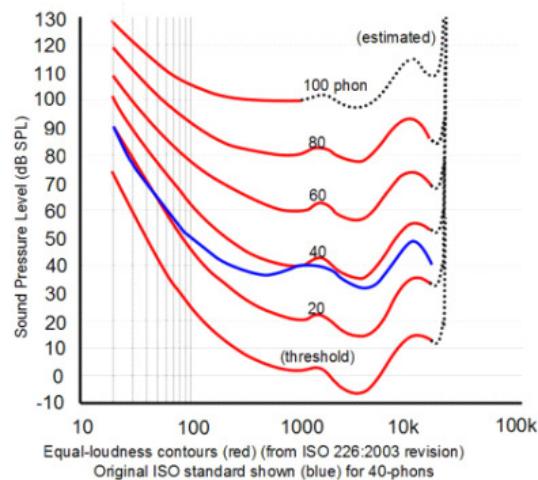
GR

Question 223.4.16

Sound	Sound Level (dB)
Jet Airplane at 30m	140
Rock Concert	120
Siren at 30m	100
Car interior when Traveling 60 mi/h	90
Street Traffic	70
Talk at 30 cm	65
Whisper	20
Rustle of Leaves	10
Quietest thing we can hear (I_o)	0

Frequency Range
Demo

Loudness and frequency



Robinson-Dadson equal loudness curves (Image in the Public Domain courtesy Lindosland)

Our ears are truly amazing in their range and ability. But, sounds with the same intensity at different frequencies do not appear to us to have the same loudness. The

frequency response graph above show how this relationship works for test subjects. We don't hear high or low frequencies as well. We have a peak response around 4000 Hz.

5 Doppler Effect and Superposition

Fundamental Concepts

- The frequency of a wave depends on the relative motion between the source and detector.
- Two waves in the same medium add up point for point at every location in the medium. This process is called superposition.

Doppler Effect

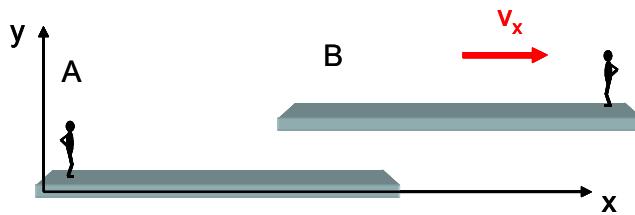
We have learned what happens when a sound wave is generated. But so far, we have assumed that sound emitter was staying still. But we know of many sound emitters that move. What happens if the emitter of the sound is moving? Worse, Back in PH121 or Dynamics we considered the relative motion between two reference frames. What happens to the sound emitter is stationary, but we, the listener, are moving?

Question 223.5.1

Doppler Ball Demo

Let's start by considering an inertial reference frame (remember this from Dynamics/PH121?)

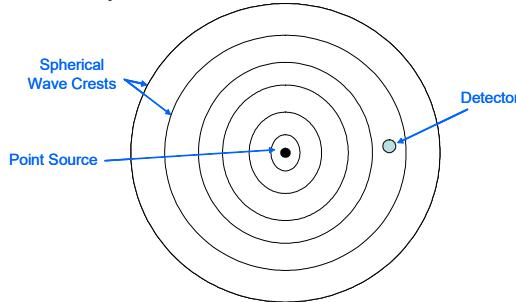
Suppose we pick two such inertial reference frames, one traveling with a velocity v_x with respect to the other. Let's also place them far away from any other object.



Person *A* sees himself as stationary and sees person *B* traveling with velocity v_x . Person *B* sees himself as stationary, and person *A* traveling with velocity $-v_x$. We can't tell which view point is correct. In fact, both are equally valid. So it seems that whether the emitter moves, or the detector moves, either way if the motion matters, it should matter the same for both cases. From this brief review, it seems that is the *relative speed* v_x that we must consider when thinking about our sound waves.

Now suppose we have a wave generator (a point source) creating spherical waves. Let the point source be at rest, say, in frame *A*.

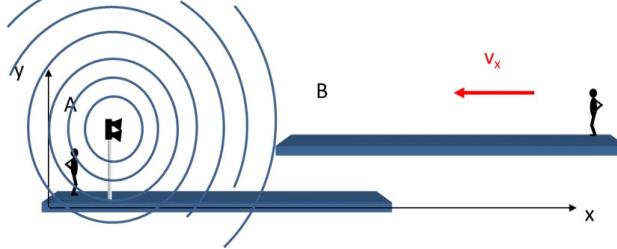
BYU Demo



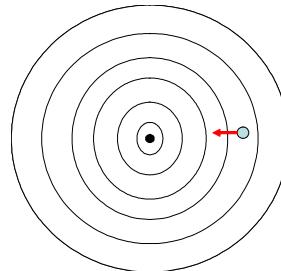
Question 223.5.2

Let's also assume a detector. If the detector is stationary with respect to the emitter, the detector sees a frequency of the wave, f just as it is created by the emitter. But let's have the detector be in frame *B*

Move George



so that it moves relative to the emitter and take our point of view from the detector frame. A top view might look like this

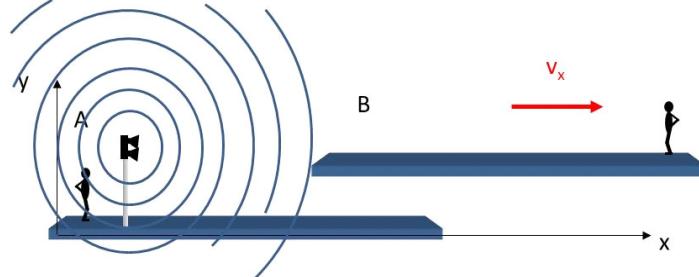


Remember, that the frequency is the number of crests that pass by a given point in a unit

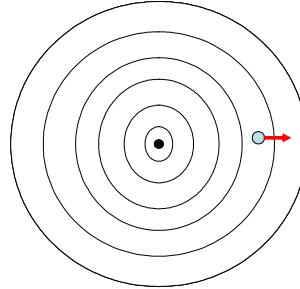
time. Does the moving detector see the same number of crests per unit time as when it was stationary?

No, the frequency appears to be higher! That is because every time a wave crest hits the detector, the detector moves toward the next wave crest.

How about if we let the detector move the other way?



The top view might look like this



Again the frequency seen by the detector is different, but this time lower. Each time a wave crest hits the detector, the detector moves away from the next wave crest, giving more time for the wave to catch up to the detector (if the period between wave crests goes up, then the frequency must go down because $f = \frac{1}{T}$).

Question 223.5.3

We can quantify this change. Take our usual variables f_A, λ_A for the stationary emitter, f_B, λ_B for the moving detector, and the velocity of sound v_{sound} . When the detector moves toward the source, it sees a different velocity. If you used the subscript system to do relative motion you might identify the speed of the sound wave created by the source in frame A as v_{aA} and the relative speed of the frame A as viewed from frame B as v_{AB} . Then the speed of sound from the source in frame A as viewed in frame B would be.

$$v_{aB} = v_{aA} + v_{AB}$$

This is our normal Galilean transformation. We could abbreviate the subscripts as just.

$$v_B = v_{\text{sound}} + v_x$$

Since the detector is riding along with frame B which is moving with speed v_x we could write

$$v_d = v_x$$

$$v_B = v_{\text{sound}} + v_d \quad (5.1)$$

where we are using v_d as the detector speed. In effect, the relative speed adds to the speed of sound making the wave crests come faster from the point of view of the detector. The wavelength will not be changed ($\lambda_A = \lambda_B$), since the distance between wave crests does not change, so

$$v = \lambda f$$

tells us the frequency must change. The new frequency f_B is given by

$$f_B = \frac{v_B}{\lambda} = \frac{v_{\text{sound}} + v_d}{\lambda}$$

We can eliminate λ from this expression for the change in f by using $v_A = \lambda f_A$ again, this time solving for λ

$$f_B = \frac{v_{\text{sound}} + v_d}{v_{\text{sound}}} f_A \quad \text{observer moving toward the source} \quad (5.2)$$

[Change the Demo](#)

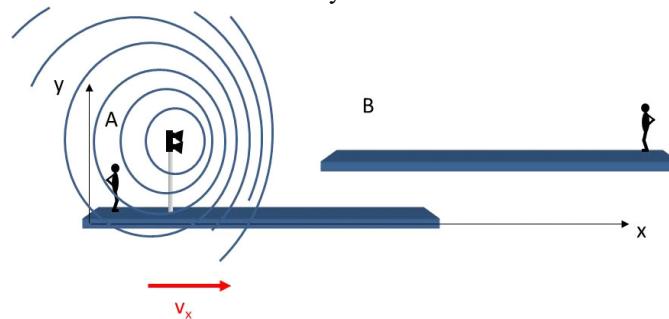
Now if the detector is going the other way v_d is negative.

$$v_B = v_{\text{sound}} - v_d$$

We expect that as the wave crest approaches the detector, the detector moves away from it. It takes longer for the crest to reach the detector. The frequency will be smaller.

$$f_B = \frac{v_{\text{sound}} - v_d}{v_{\text{sound}}} f_A \quad \text{observer moving away from the source} \quad (5.3)$$

From our thinking about the motion of two inertial reference frames, we expect a similar situation if the detector is stationary and the source moves.

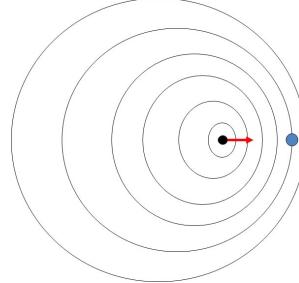


Since the emitter is now riding along with frame A at the relative speed, v_x we could

write

$$v_e = v_x$$

where v_e is the speed of the emitter. In this case the detector will see a different wavelength. The top down view might look like this



In fact, if we measure the distance between the crests we must account for the fact that the source moved by an amount

$$\Delta x = v_e T = \frac{v_e}{f_A}$$

during one period. Then the wavelength is seen to be shorter by this amount.

$$\lambda_B = \lambda_A - \frac{v_e}{f_A}$$

Using the basic equation

$$\lambda = \frac{v}{f}$$

once more, we can write the frequency as

$$f_B = \frac{v_{sound}}{\lambda_B} = \frac{v_{sound}}{\lambda_A - \frac{v_e}{f_A}}$$

and once more using the basic equation

$$\lambda = \frac{v}{f}$$

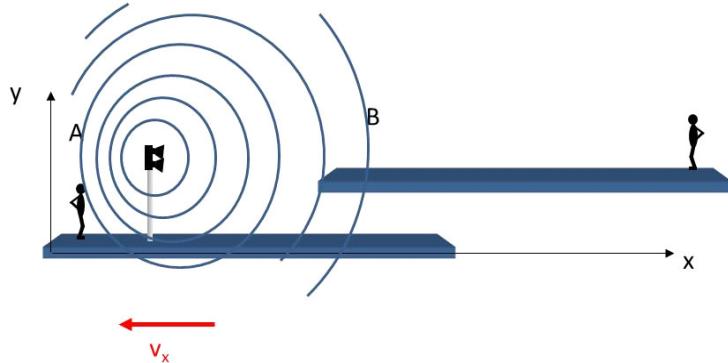
we can write this as

$$f_B = \frac{v_{sound}}{\frac{v_{sound}}{f_A} - \frac{v_e}{f_A}}$$

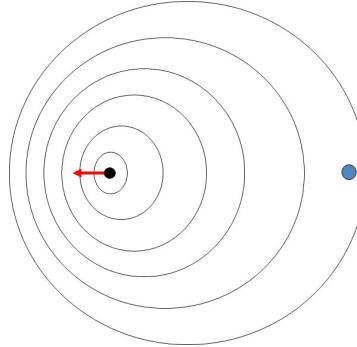
or

$$f_B = \frac{v_{sound}}{v_{sound} - v_e} f_A \quad \text{Source moving toward observer} \quad (5.4)$$

When the source is moving away from the detector,



the top down view might look like this



we expect the wavelength to be larger. This gives

$$f_B = \frac{v_{sound}}{v_{sound} + v_e} f_A \quad \text{Source moving away from observer} \quad (5.5)$$

We can combine these formulae to make one expression

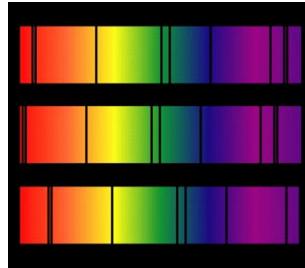
$$f_B = \frac{v_{sound} \pm v_d}{v_{sound} \mp v_e} f_A \quad (5.6)$$

where we use the top sign for the speed when the mover (detector or emitter) is going toward the non-mover.

We can see that by moving the emitter or the detector, we get a frequency change. This fact is named after the scientist who studied it. It is called the *Doppler effect* and the change in frequency is called the *Doppler shift*.

Doppler effect in light

Light is also a wave, and so we would expect a Doppler shift in light. Indeed we do see a Doppler shift when we look at moving glowing objects. Here is an optical spectrum of the Sun on the top and a spectrum of a similar star moving away from us in the middle. The final spectrum is for a star moving toward us.



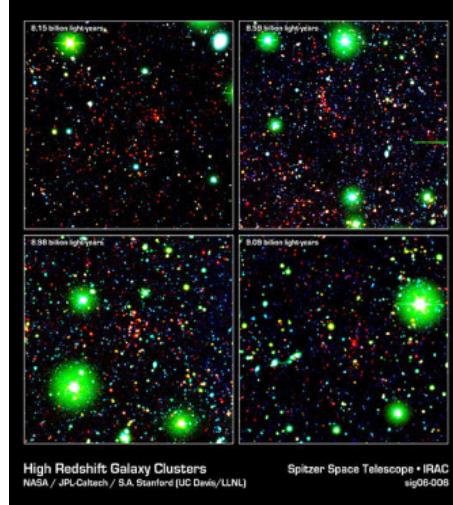
Top: Normal 'dark' spectral line positions at rest. Middle: Source moving away from observer. Bottom: Source moving towards observer. (Public domain image courtesy

NASA: <http://www.jwst.nasa.gov/education/7Page45.pdf>)

Note that the wavelength of the lines is shifted toward the red part of the spectrum when the glowing object moves away from us. This is equivalent to lowering of the frequency of a truck engine noise as it goes away from us. The larger wavelengths indicate a lower frequency of light because

$$f = \frac{c}{\lambda}$$

This gives us a way to determine if distant stars and galaxies are moving toward or away from us. We look for the chemical signature pattern of lines, then see whether they are shifted to the red (moving away from us) or blue (moving toward us) compared to the position in their spectrum of the Sun. This photo is of some of the most distant galaxies that are moving very fast away from us. Their redshift is very large.



High Redshift Galaxy Cluster shown here in false color from the Spitzer Space Telescope. (Public domain image courtesy NASA/JPL-Caltech/S.A. Stanford (UC Davis/LLNL)

Deriving the Doppler equation for light is more tricky because the speed of light is constant in all reference frames. We really tackle this in our PH279 class. So I will just quote the result here.

$$\lambda_- = \lambda_o \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}} \quad \text{receding source} \quad (5.7)$$

$$\lambda_- = \lambda_o \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}} \quad \text{Approaching source} \quad (5.8)$$

Superposition Principle

Wave Machine

Demo

Question 223.5.5

What happens if we have more than one wave propagating in a medium? You probably experienced this as a child. Your parents made you take a bath. You discovered that you could make waves with your arm. But chances are you have two arms, and that you discovered you could make two waves, one with each arm. And when the two waves met in the middle, the water left the bath tub! What happened was that the two wave crests met in the same place and the medium (water) piled up there. We call the combination of two waves in the same medium *superposition*. The word literally means putting one wave on top of another. When we superimpose two waves, their wave functions simply add.

Definition 5.1 *Superposition: If two or more traveling waves are moving through a medium, the resultant wave formed at any point is the algebraic sum of the values of the individual wave forms.*

So if we have

$$y_1(x, t) \quad (5.9)$$

and

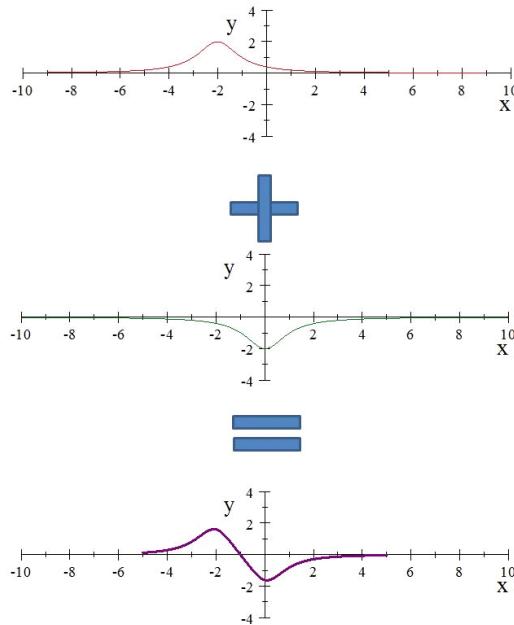
$$y_2(x, t) \quad (5.10)$$

both propagating on a string, then we would see a resultant wave

$$y_r(x, t) = y_1(x, t) + y_2(x, t) \quad (5.11)$$

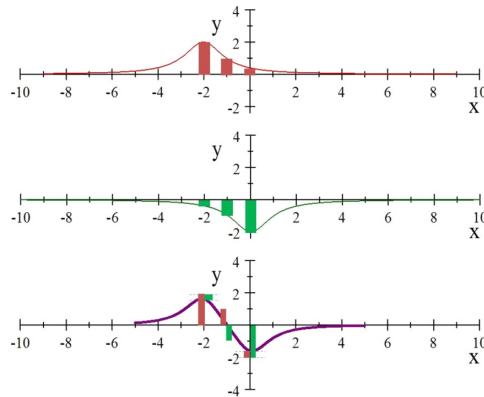
This is a fantastically simple way for the universe to act!

Let's look at an example. let's add the top wave (red) to the middle wave (green). We get the bottom wave (purple)



Of course we are adding these in the snapshot view. So this is all done for just one instant of time.

Let's see how to do this.



As an example, start at $x = -2$. In the figure, I drew a red bar to show the y value at $x = -2$ for the red curve. Likewise, I have a green bar showing the value of y at $x = -2$ for the green wave. Note that this is negative. On the bottom graph, the bars have been repeated, and we can see that the red bar minus the green bar brings us to the value for the resulting wave at the point $x = -2$. We need to do this at every point along all the waves for this instant of time.

This is tedious by hand, so we won't generally do this calculation by hand. But a computer can do it easily.

Note that this is really only true for *linear* systems. Let's take the example of a Slinky™. If we form two waves in the Slinky, they behave according to the superposition principle most of the time. But suppose we make the amplitude of the individual waves large. They may travel individually OK, but when the amplitudes add we may overstretch the Slinky. Then it would never return to its original shape. The wave form would have to change. Such a wave is not linear. There is a good rule of thumb for when waves are linear.

A wave is generally linear when its amplitude is much smaller than its wavelength.

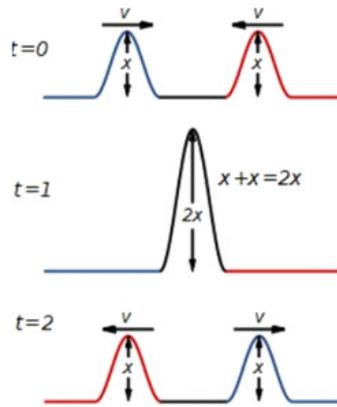
Consequences of superposition

Question 223.5.6

Question 223.5.7

Question 223.5.8

Linear waves traveling in media can pass through each other without being destroyed or altered!



Constructive Interference (Public Domain image by Inductiveload,
http://commons.wikimedia.org/wiki/File:Constructive_interference.svg)

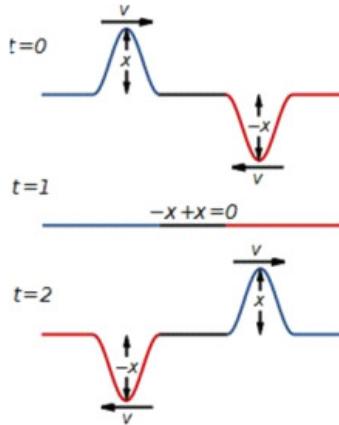
Our wave on the string makes the string segments move in the y -direction. Both waves do this. So putting the two waves together just makes the string segments move more! There is a special name for what we observe

Definition 5.2 interference: *The combination of separate waves in the same region of space to produce a resultant wave.*

We also have a special name for when the amplitude of the resultant wave gets larger.

Definition 5.3 Constructive Interference: *interference between waves when the displacements caused by the two waves are in the same direction*

What happens if one of the pulses is inverted?



Destructive Interference (Public Domain image by Inductiveload,
http://commons.wikimedia.org/wiki/File:Destructive_interference1.svg)

When the two pulses meet, they “cancel each other out.” But do they go away? No! the energy is still there, the string segment motions have just summed vectorially to zero, the energy carried by each wave is still there in the stretched string. Because we momentarily seem to destroy the wave pulses, we call this type of interference “destructive interference.”

Definition 5.4 Destructive Interference: *Interference between waves when the displacements caused by the two waves are opposite in direction*

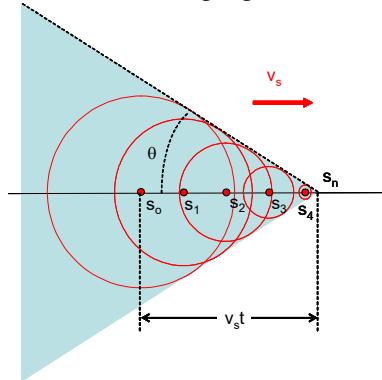
Superposition and Doppler: Shock waves

Question 223.5.4

What happens when the speed of the source is greater than the wave speed?

Remember that the wave speed depends only on the medium. Let's call the crests of a wave the *wave front*. In the picture below, a point source is generating a wave and the red lines are the wave fronts.

When $v_s = v_{sound}$ the waves superimpose. They begin to pile up. If we allow $v_s > v_{sound}$ then the wave fronts are no longer generated within each other.



The leading edge of the wave fronts superimpose to form a cone shape. The half angle of this cone is called the *Mach angle*

$$\sin \theta = \frac{vt}{v_s t} = \frac{v}{v_s} \quad (5.12)$$

This ratio v/v_s is called the Mach number and the conical wave front is called a shock wave. We see them often in water

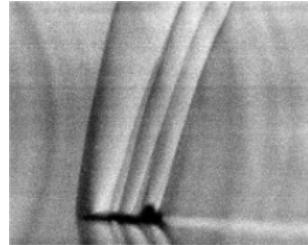


Boat wakes as a Doppler cone. Image courtesy US Navy. Image is in the Public Domain.

Doppler Movie

and hear them when jet aircraft go supersonic. In the next figure we can see a picture of a T-38 breaking the sound barrier. You can see the Mach cones, but notice that there are several! Remember that a disturbance creates a wave. There are disturbances cre-

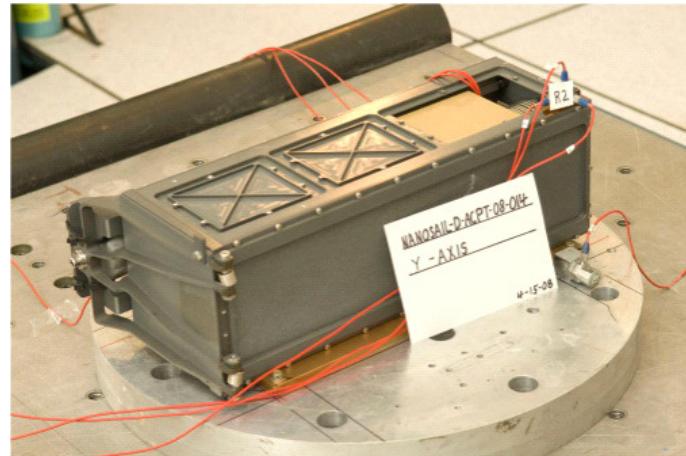
ated by the nose of the plane, the rudder, and the wings, and perhaps the cockpit in this Schlieren photograph.



Dr. Leonard Weinstein's Schlieren photograph of a T-38 Talon at Mach 1.1, altitude 13,700 feet, taken at NASA Langley Research Center, Wallops in 1993. Image Courtesy NASA, image is in the Public Domain.

Importance of superposition

The combination of waves is important for both scientists and engineers. In engineering this is the heart of vibrometry.



Marshall and Cal Poly technicians wired the NanoSail-D spacecraft to accelerometers, instruments which measure vibration response during simulated launch conditions.

Image courtesy NASA, image in the Public Domain.

Mechanical systems have moving parts. These moving parts can be the disturbance that creates a wave. If more than one wave crest arrives at a location in the device, the amplitude at that location could become large. The oscillation of this part of the device could rattle apart welds or bolts, destroying the device. Later, as we study spectroscopy, we will see how to diagnose such a problem and hint at how to correct it.

6 Standing Waves

A special case of superposition is that of two waves of the same frequency traveling opposite directions. Mixing two such waves can give rise to resonant patterns. These resonant patterns are the basis of music, and are of concern in building structures, among other things.

Fundamental Concepts

- When a wave meets a boundary, it will reflect
- Reflected waves will invert if the boundary is fixed or more like a fixed boundary.
- Reflected waves will not invert if the boundary is free or more like a free boundary.
- Two waves of equal frequency but traveling opposite directions can cause resonant patterns called standing waves.
- Only certain frequencies will produce standing waves. The boundary conditions determine which frequencies will work.
- The series of frequencies that produce standing waves is called the harmonic series.

Mathematical Description of Superposition

We know what superposition is, but we don't really want to add values for millions of points in a medium to find out what a combination of waves will look like. At the very least, we want to make a computer do that (and programs like OpenFoam do something very akin to this!). But where we can, we would like to combine wave functions algebraically. Let's see how this can work.

Let's define two wave functions

$$y_1 = A \sin(kx - \omega t)$$

and

$$y_2 = A \sin(kx - \omega t + \phi_o)$$

These are two waves with the same frequency and wave number traveling the same direction in the medium, but they start at different times. The graph of y_2 is shifted by

an amount ϕ_o .

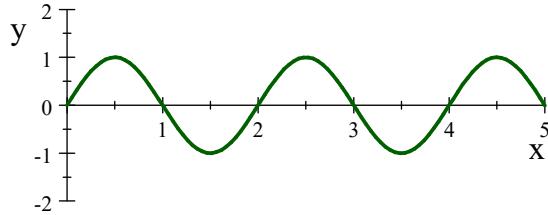
I will pick some values for the constants

$$\begin{aligned}\lambda &= 2 \\ k &= \frac{2\pi}{\lambda} \\ \omega &= 1 \\ \phi_o &= \frac{\pi}{6} \\ t &= 0 \\ A &= 1\end{aligned}$$

then for y_1 we have

$$\begin{aligned}y_1 &= (1) \sin \left(\frac{2\pi}{\lambda} x - (1)t \right) \\ &= \sin \left(\frac{2\pi}{2} x - (1)t \right) \\ &= \sin (\pi x - t)\end{aligned}$$

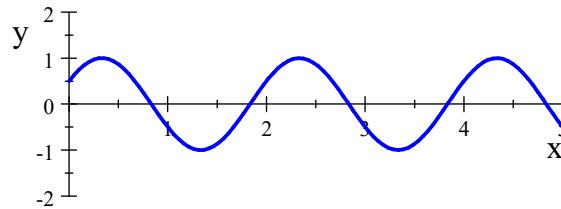
here is a plot of the wave function, y_1



Now let's consider y_2 . Using the values we chose, y_2 can be written as

$$\begin{aligned}y_2 &= A \sin (kx - \omega t + \phi_o) \\ &= \sin \left(\pi x - t + \frac{\pi}{6} \right)\end{aligned}$$

which looks like this



What does it look like if we add these waves using superposition? Symbolically we

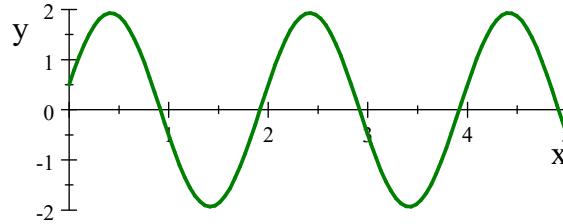
have

$$y_r = A \sin(kx - \omega t) + A \sin(kx - \omega t + \phi_o) \quad (6.1)$$

and putting in the numbers gives

$$y_r = \sin(\pi x - t) + \sin\left(\pi x - t + \frac{\pi}{6}\right)$$

which is shown in the next graph.



Notice that the wave form is taller (larger amplitude). Also notice it is shifted along the x axis.

We can find out how much by rewriting y_r . We want to rewrite equation (6.1) so it is easier to interpret. To do this we need to remember a trig identity

$$\sin a + \sin b = 2 \cos\left(\frac{a-b}{2}\right) \sin\left(\frac{a+b}{2}\right)$$

Then, for our case, let $a = kx - \omega t$ and $b = kx - \omega t + \phi_o$. This lets us rewrite our resultant wave.

$$\begin{aligned} y_r &= A \sin(kx - \omega t) + A \sin(kx - \omega t + \phi_o) \\ &= 2A \cos\left(\frac{(kx - \omega t) - (kx - \omega t + \phi_o)}{2}\right) \sin\left(\frac{(kx - \omega t) + (kx - \omega t + \phi_o)}{2}\right) \\ &= 2A \cos\left(\frac{-\phi_o}{2}\right) \sin\left(\frac{2kx - 2\omega t + \phi_o}{2}\right) \\ &= 2A \cos\left(\frac{-\phi_o}{2}\right) \sin\left(kx - \omega t + \frac{\phi_o}{2}\right) \\ &= 2A \cos\left(\frac{\phi_o}{2}\right) \sin\left(kx - \omega t + \frac{\phi_o}{2}\right) \end{aligned}$$

where we used the fact that $\cos(-\theta) = \cos(\theta)$.

To interpret this new form of our resultant wave equation, let's look at the parts of this expression. First take

$$\sin\left(kx - \omega t + \frac{\phi_o}{2}\right) \quad (6.2)$$

This part is a traveling wave with the same k and ω as our original waves, but it has a

phase constant of $\phi_o/2$. So our combined wave is shifted by $\phi_o/2$ or half the phase shift of y_2 .

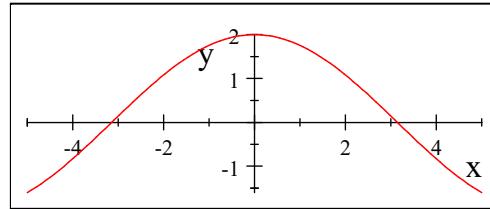
Now let's look at other factor

$$2A \cos\left(\frac{\phi_o}{2}\right) \quad (6.3)$$

This part has no time dependence. We recognize from our basic equation

$$y(x, t) = A \sin(kx - \omega t + \phi_o)$$

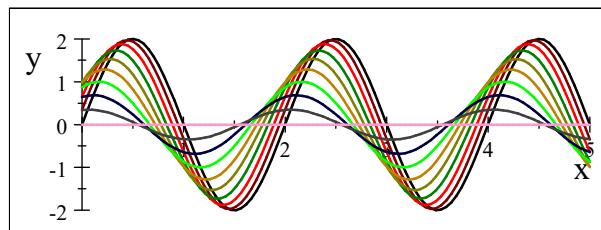
that the amplitude, A is a constant—not dependent on x or t , that multiplies the sine function. But now we have a more complex term that is not dependent on x or t that multiplies the sine function. The whole term must be the new amplitude! It has a maximum value when $\phi_o = 0$



When $\phi_o = \pi$, then

$$2A \cos\left(\frac{\pi}{2}\right) = 0$$

so when $\phi_o = 0$ we have a new maximum amplitude of twice the original amplitude, $2A$, and when $\phi_o = \pi$ we have no amplitude. Here is our wave for several choices of ϕ_o .



We can see that in our case the fact that the two waves added to produce a larger amplitude was just luck. We could have gotten anything from twice the single wave amplitude to no amplitude at all.

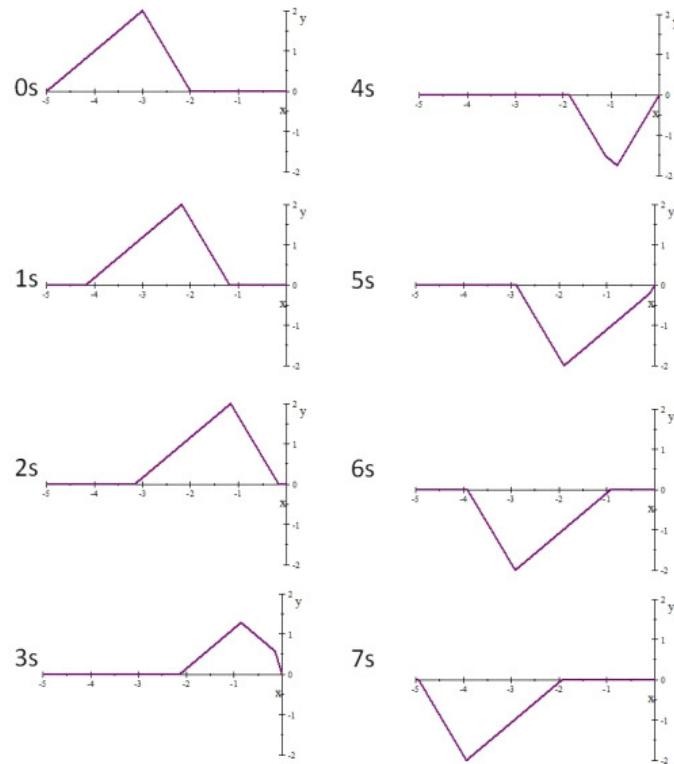
Reflection and Transmission

In our examples so far, we have not explained how we got two waves into a medium. One way is to simply reflect one wave back on top of itself.

Spring Pulse Inversion Demo
In class we made pulses on a long spring with one end of the spring fixed (held by a class member). What happens when the pulse reaches the end of the rope?

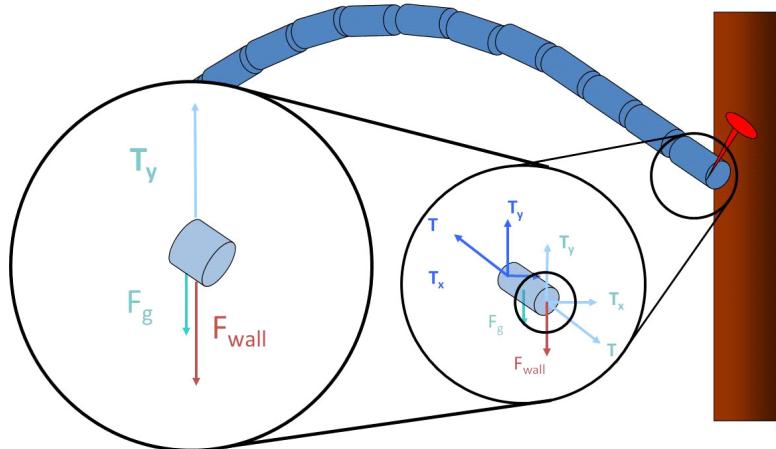
Case I: Fixed rope end.

Question 223.6.1



There is a big change in the medium at the end of the rope—the rope ends. This change in medium causes a reflection.

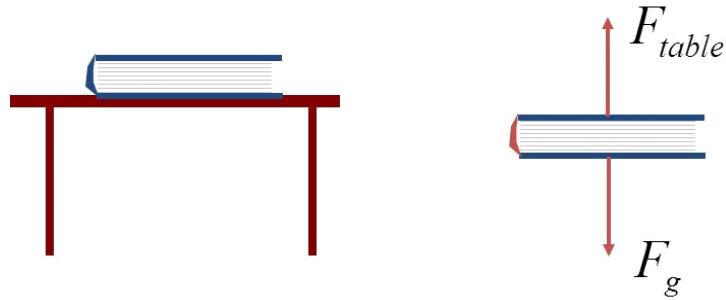
In the fixed end case, the pulse is inverted. We should consider why this inversion happens.



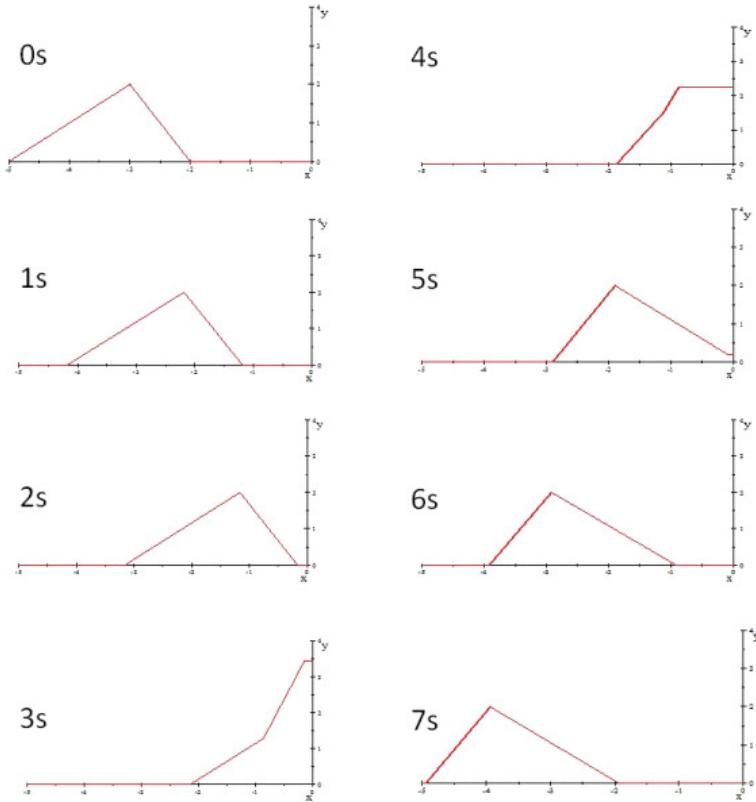
The end of the rope pushes up on the support (person, or nail or whatever). By Newton's third law the support must push back in an equal, but opposite direction, on the rope. This force sends the pieces of rope near it downward. We could think of the squashed nail atoms as having been given an amount of spring potential energy. They will transfer this energy back to the rope by pushing the end of the rope down. This downward motion becomes a new pulse that is an inversion of the original pulse traveling the opposite direction.

Popper demo

Does this seem reasonable? Remember studying normal forces? Consider a book on a table. The book has a force due to gravity. The table exerts a force equal to $m_{book}g$ on the book, or else the book smashes through the table.

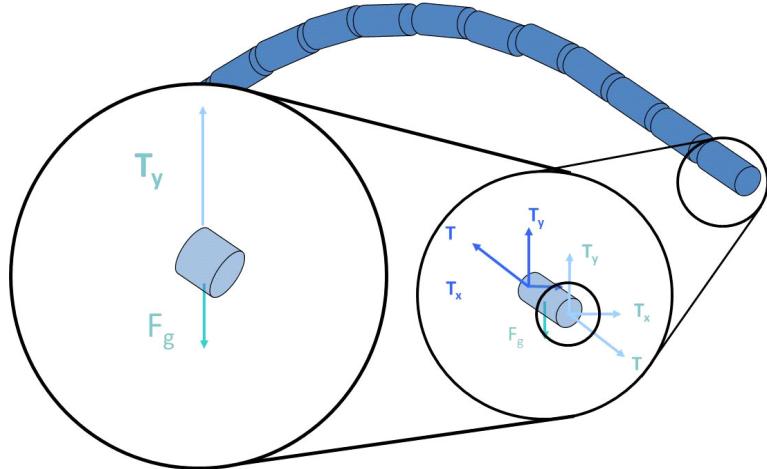


The normal force exerted by the atoms of the table keep the book up. It is this same type of force that keeps the rope on the nail. The atoms of the nail must push down on the end of the rope. They exert a force and this causes the inversion.

Case II: Loose rope end.

But what happens if the rope end is not fixed?

The rope end rises, and therefore there is no force exerted. The pulse (or at least part of the pulse energy) is still reflected, but there is no inversion!



The end of the rope will come down, but the reason is that the force due to gravity acts on the mass of the rope end. The energy of the wave was made into potential energy of the rope end. As the rope end loses potential energy, that energy is put back into a wave going the opposite direction.

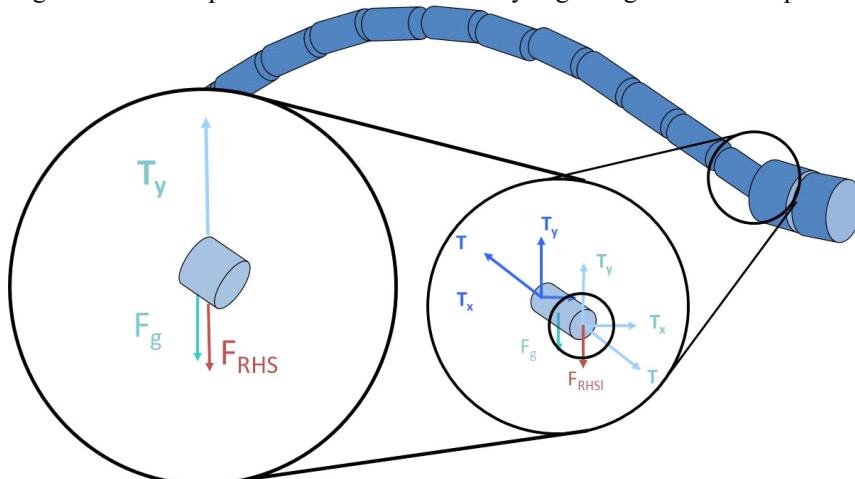
Case III: Partially attached rope end

Question 223.6.2

Question 223.6.3

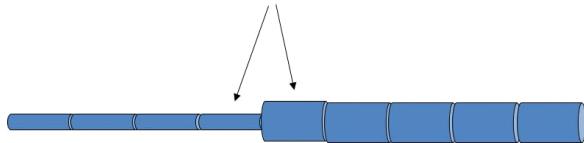
Now lets tie the rope to another rope that is larger, more dense, than the rope we have been using, what will happen when we make waves in this combined rope?

The light end of the rope exerts a force on the heavy beginning of the new rope

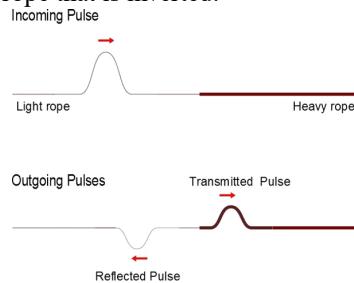


In this case consider momentum

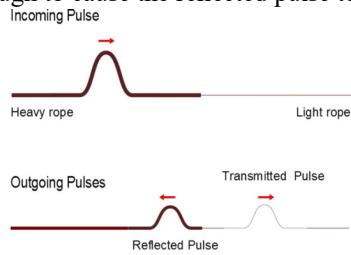
Which piece is harder to move?



The heavier rope resists being moved because of its larger mass. This resistance to motion is a little like our fixed end case. It is harder to transfer energy to the heavy rope, and the heavy rope resists the pull of the light rope. This resistance pushes back on the end of the light rope. This is a downward push. So once again we will have a reflected pulse in the light rope that is inverted.



We could also make a pulse in the heavy rope. What would happen then when the pulse reached the interface? You might be able to guess that the light rope won't have much effect on the end of the heavy rope. The light rope will cause a reflection, but it's weak downward push is not enough to cause the reflected pulse to invert.



Going from a heavy rope to a light rope makes an interface that is more like a free end.

Notice that in both cases there is a *transmitted* pulse. The transmitted pulse is what is left of the energy from the original pulse that has not been reflected. So we would not expect it to be inverted, and, indeed, it never is. We have split the amount of energy traveling along the rope

Mathematical description of standing waves

Standing Wave Demo

Question 223.6.4

Now that we have a way to make two waves to superimpose, we can study the special case of a reflected wave. We will find that this special case can produce interesting patterns of constructive and destructive interference.

The patterns of constructive and destructive interference are the result of the superposition of two traveling waves with the same frequency going in opposite directions. Let's start with two standing waves with the same phase constant for simplicity.

$$\begin{aligned}y_1 &= A \sin(kx - \omega t) \\y_2 &= A \sin(kx + \omega t)\end{aligned}$$

The sum is

$$y = y_1 + y_2 = A \sin(kx - \omega t) + A \sin(kx + \omega t)$$

To gain insight into what these two waves produce, we use another of our favorite trig identities

$$\sin(a \pm b) = \sin(a) \cos(b) \pm \cos(a) \sin(b)$$

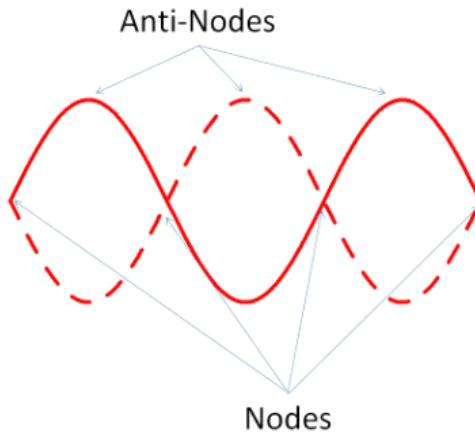
to get

$$\begin{aligned}y &= A \sin(kx - \omega t) + A \sin(kx + \omega t) \\&= A \sin(kx) \cos(\omega t) - A \cos(kx) \sin(\omega t) + A \sin(kx) \cos(\omega t) + A \cos(kx) \sin(\omega t) \\&= 2A \sin(kx) \cos(\omega t) \\&= (2A \sin(kx)) \cos(\omega t)\end{aligned}$$

This looks like the harmonic oscillator equation

$$y = A \cos(\omega t + \phi_o)$$

with $\phi_o = 0$. The factor $2A \sin(kx)$ has no time dependence, so it could be considered the amplitude of the harmonic oscillator. But this is a very odd amplitude. It depends on position. That is, we can view the rope as a set of harmonic oscillators who's amplitudes are different for each value of x .



Question 223.6.5

Question 223.6.6

But this is just what we see in our standing wave! We can identify spots along the x axis where the amplitude is always zero! we will call these spots *nodes*. These happen when $\sin(kx) = 0$ or when

$$kx = n\pi$$

By using

$$k = \frac{2\pi}{\lambda}$$

we have

$$\begin{aligned} \frac{2\pi}{\lambda}x &= n\pi \\ \frac{2}{\lambda}x &= n \\ x &= n\frac{\lambda}{2} \end{aligned}$$

We can also find the places along x where the amplitude will be largest. this occurs when $\sin(kx) = 1$ or when

$$kx = n\frac{\pi}{2}$$

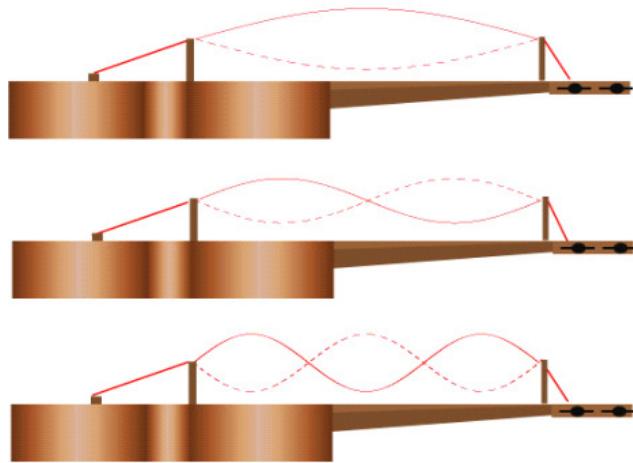
or

$$\begin{aligned} \frac{2\pi}{\lambda}x &= n\frac{\pi}{2} \\ x &= n\frac{\lambda}{4} \end{aligned}$$

these are called *antinodes*.

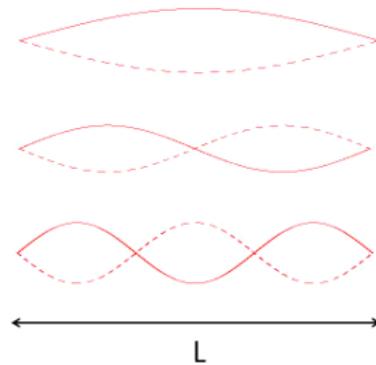
This combination of two waves does not look like it goes anywhere. It seems to “stand” in place. We call it a *standing wave*. We can also create standing waves with sound or even light waves! But let’s look at standing waves in some detail first.

Standing Waves in a String Fixed at Both Ends



Question 223.6.7

If we attach a string to something on both ends, we find something interesting in the standing wave pattern. Not all imaginable standing waves can be realized. Some frequencies are preferred, and some never show up. These non-preferred frequencies will make waves, but not standing waves. We say that the standing wave pattern is *quantized*, meaning that only certain frequency values will make a standing wave pattern. The patterns that are allowed are called *normal modes*. We will see this any time a wave confined by boundary conditions (light in a resonant cavity, radio waves in a wave guide, electrons in an atom, etc.). In the last figure we saw some standing waves on a ukulele string. But we can draw the standing waves without the instrument.



The figure shows three normal modes for a string. Of course there are many more.

We find which modes are allowed by first imposing the boundary condition that each end must be a node. We start with

$$y = 2A \sin(kx) \cos(\omega t)$$

and recognize that we have one condition met because $y = 0$ when $x = 0$. We need $y = 0$ when $x = L$. That happens when

$$kL = n\pi$$

I will write this as

$$k_n L = n\pi$$

to indicate there are many values of k that could make a standing wave pattern. Solving this for λ_n gives

$$\begin{aligned} \frac{2\pi}{\lambda_n} L &= n\pi \\ \frac{2L}{n} &= \lambda_n \end{aligned}$$

Let's see how this works, the first mode will have

$$\lambda_1 = 2L$$

where L is the length of the string. Looking at the figure, we can see that this is true. The first normal mode has a length that is half the first mode wavelength.

The second mode has three nodes (one on each end and one in the middle). This gives

$$\lambda_2 = L$$

We can keep going, the third mode will have five nodes

$$\lambda_3 = \frac{2L}{3}$$

and so forth to give

$$\lambda_n = \frac{2L}{n}$$

We use our old friend

$$v = f\lambda$$

to find the frequencies of the modes

$$f = \frac{v}{\lambda}$$

Thus

$$f_1 = \frac{v}{\lambda_1} = \frac{v}{2L}$$

or, in general

$$\begin{aligned} f_n &= \frac{v}{\lambda_n} = n \frac{v}{2L} \\ &= \frac{n}{2L} v \\ &= \frac{n}{2L} \sqrt{\frac{T}{\mu}} \end{aligned}$$

The lowest frequency that works has a special name, the *fundamental frequency*. The higher frequencies are integer multiples of the fundamental. When this happens we say that the frequencies form a *harmonic series*, and the modes are called *harmonics*.

Question 223.6.7

Question 223.6.9

Starting the waves

So, suppose we do not have a vibrating blade to make the wave patterns for a string that is fixed at both ends. Can we just pluck it to make it vibrate on a natural frequency?

Yes! only the normal modes will be excited by the pluck, any other frequencies will die out quickly (we won't show this mathematically in this class). So the only allowed frequencies (the ones that will result from a pluck) are the natural frequencies or harmonics. The frequency of waves on the string is *quantized!* That is, only some values are allowed. This idea is the basis behind Quantum mechanics (which views light and even matter as waves).

Musical Strings

So how do we get different notes on a guitar or Piano?

$$f_n = \frac{n}{2L} \sqrt{\frac{T}{\mu}} \quad (6.4)$$

A guitar uses tension to change the frequency or pitch (tuning) and length of string (your fingers pressing on the strings) to change notes.

A Piano uses both tension and length of string (and mass per unit length as well!). What do you expect an organ will do?

7 Light and Sound Standing waves

Reading Assignment 21.4, 21.5

Fundamental Concepts

- The harmonic series expressed by a system experiencing standing waves depends on the *boundary conditions*.
- The harmonic series for open pipes is different than the harmonic series for a pipe closed on one end.
- Energy persists in the waves that have the harmonic series frequencies because of resonance.

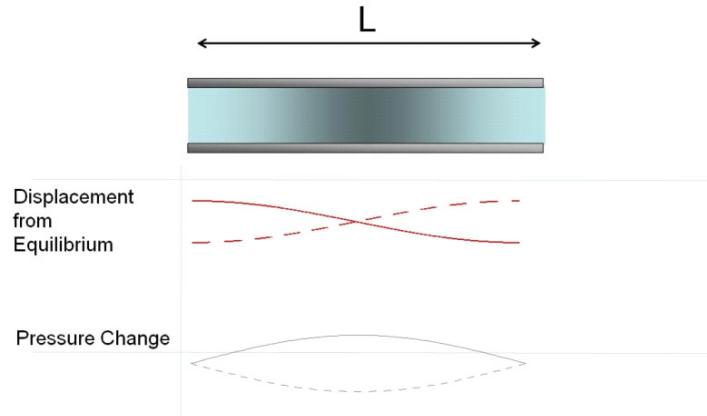
Sound Standing waves (music)

Suppose we send a sound wave down a pipe. When the air molecules strike the molecules next to them they end up being reflected back. This happens as the wave goes down the pipe until the wave reaches the end of the pipe. Remember that where the molecules bunch up, the pressure is higher.

Question 223.7.1

When we reach the end of the pipe the molecules can't bounce off the walls of the pipe anymore. They travel out into the surrounding air. It is harder to get the room pressure to change because molecules can come from any direction to fill up a vacancy. This is an effective medium change, and there will be some energy reflected back from this pipe-to-room interface. The reflected wave can make a standing wave. This is the basis of wind instruments. Let's repeat the analysis we did last time and find the possible frequencies that can make a standing wave, but this time for a sound wave in a pipe.

Take a pipe as shown in the next figures.



If we have a pipe open at both ends, we can see that air molecules are free to move in and out of the ends of the pipe. If the air molecules can move, the ends must not be nodes. This is different than the string case we studied last time! We expect that there must be a node somewhere. We can reasonably guess that there will be a node in the middle of the pipe due to symmetry. Of course, the pressure on both ends must be atmospheric pressure. So, remembering that pressure and displacement are 90° out of phase for sound waves, we can guess that there are pressure nodes on both ends.

For the first harmonic we can draw a displacement node in the middle and we see that

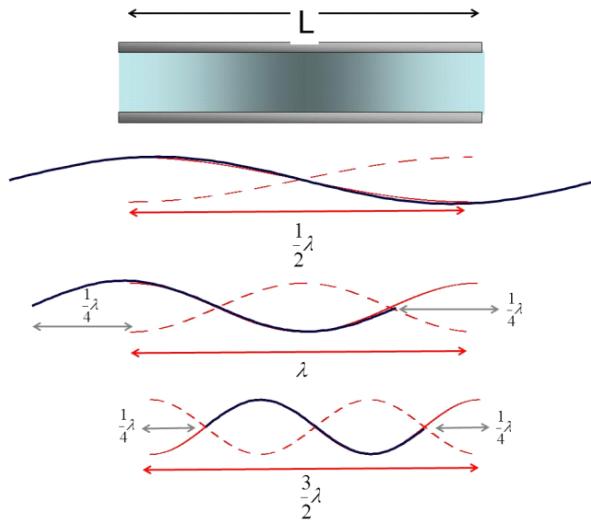
$$\lambda_1 = 2L$$

It takes two lengths of pipe to have the same length as the wavelength that is in our one single pipe. Of course, our wave is hanging out of our pipe. But if we set two additional pipes of length L along side our pipe, these two pipes would be the same length as the wavelength. The frequency would be.

$$f_1 = \frac{v}{2L} \quad (7.1)$$

The next mode fits a whole wavelength

$$\begin{aligned}\lambda_2 &= L \\ f_2 &= \frac{v}{L}\end{aligned}$$



but the next mode fits a wavelength and a half

$$\begin{aligned}\lambda_3 &= \frac{2}{3}L \\ f_3 &= \frac{3v}{2L}\end{aligned}$$

If we keep going

$$\lambda_n = \frac{2}{n}L \quad (7.2)$$

$$f_n = n \frac{v}{2L} \quad n = 1, 2, 3, 4 \dots \quad (7.3)$$

Boom Whacker and Length

This is the same mathematical form that we achieved for a standing wave on a string!

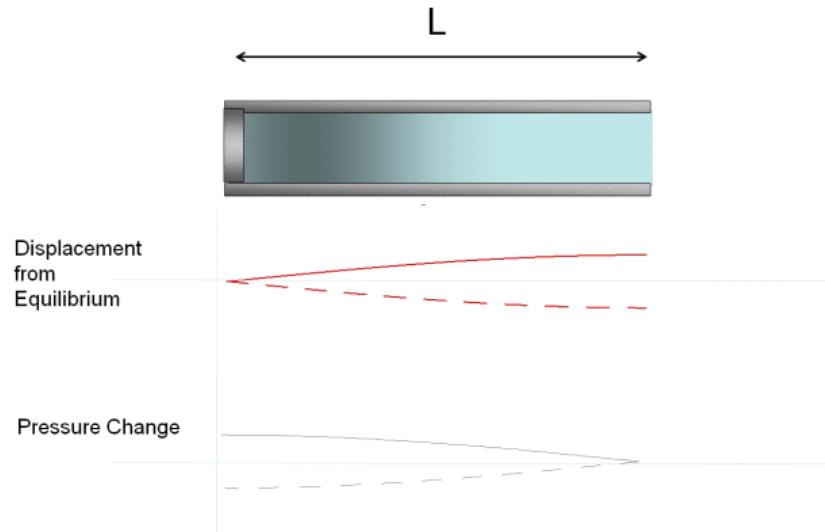
Question 223.7.2

Pipes closed on one end

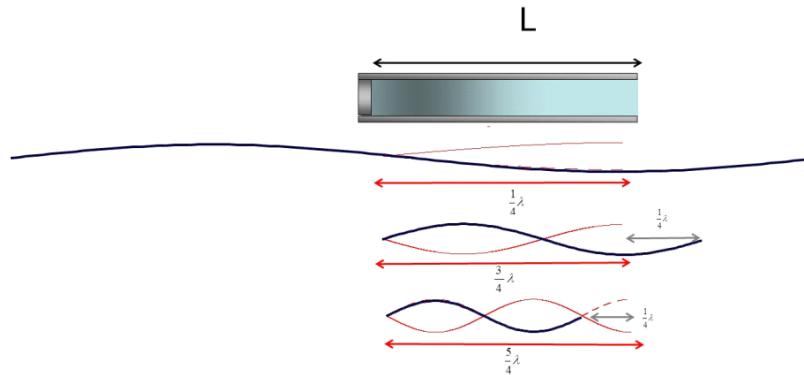
Boom Whacker Demo

But what happens if we put a cap on one end of the pipe? The air molecules cannot move longitudinally once they hit the end. This must be a displacement node. So then it must also be a pressure anti-node.

The open end is a pressure node because it stays at atmospheric pressure. This is just the same as the open ends in the open pipe case we did before. The simplest possible standing wave is shown below.



In the next figure we draw the first few harmonics for this case.



The first harmonic for the closed pipe are found by using

$$v = \lambda f$$

$$f = \frac{v}{\lambda}$$

just as we did for the string and open pipe cases. We know the speed of sound, so we have v . Knowing that the first harmonic has a node at one end and an anti node at the other end gives us the wavelength. If the pipe is L in length, then L must be

$$L = \frac{1}{4}\lambda_1$$

or

$$\lambda_1 = 4L$$

We see it now takes four lengths of pipe to be the same size as the wave that is in our

single pipe! Then the frequency is given by

$$f_1 = \frac{v}{\lambda_1} = \frac{v}{4L}$$

The next configuration that will have a node on one end and an antinode on the other will have

$$L = \frac{3}{4}\lambda_2$$

which gives

$$\lambda_2 = \frac{4}{3}L$$

and

$$f_2 = \frac{v}{\lambda_2} = \frac{3v}{4L}$$

If we continued, we would find

$$\lambda_n = \frac{4}{n}L \quad (7.4)$$

and

$$f_n = n \frac{v}{4L} \quad n = 1, 3, 5 \dots \quad (7.5)$$

This is different from the string and open pipe cases. Note that only odd values of n make a standing wave. Changing the end condition changed which frequencies would make standing waves.

Example: organ pipe



Organ Pipe Demo

The organ pipe shown is closed at one end so we expect

$$f_n = n \frac{v}{4L} \quad n = 1, 3, 5 \dots \quad (7.6)$$

Measuring the pipe, and assuming about 20 °C for the room temperature we have

$$\begin{aligned} L &= 0.41 \text{ m} \\ R &= 0.06 \text{ m} \\ v &= 343 \frac{\text{m}}{\text{s}} \end{aligned} \quad (7.7)$$

There is a detail we have ignored in our analysis, the width of the pipe matters a little. I will include a fudge factor to account for this. With the fudge factor, the wavelength is

$$\lambda_1 = 4(L + 0.6R)$$

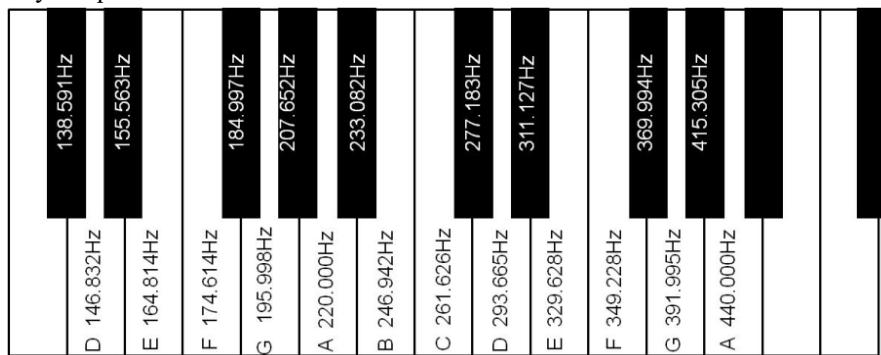
$$= 178.4 \text{ cm}$$

then our fundamental frequency is

$$f_1 = \frac{v}{\lambda_1} \quad (7.8)$$

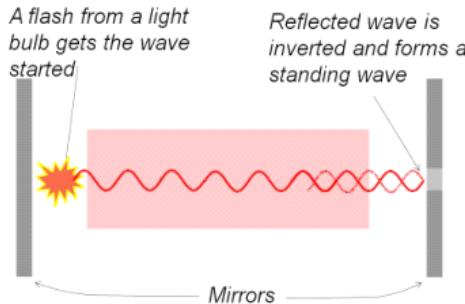
$$= 192.26 \quad (7.9)$$

We can identify this note, and compare to a standard, like a tuning fork or a piano to verify our prediction.



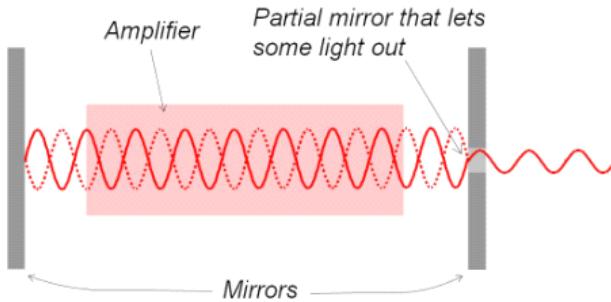
Lasers and standing waves

Light is a wave. Can we make a standing wave with light? The answer is yes, and surprisingly we do it all the time. A laser creates a standing wave as part of its amplification system. Here is a laser just getting started.



A flash of light from a flash bulb send light out in all directions. But some of the light goes to the right toward a mirror. This is the light that will eventually become our laser

beam. This light is reflected off of the mirror. The wave inverts and travels back along the center of the laser. Because it is inverted, it can cause destructive interference in places along the laser cavity. But this only works if we have just the right frequency of light. The light has to fit an integer number of half wavelengths between the two mirrors for the standing wave to form.



So only certain frequencies will work. That is why lasers usually only have one color, different frequencies of light give us different colors. So a red laser has a frequency of about 4.762×10^{14} Hz. We would expect another frequency to work that is twice this fundamental frequency.

$$\begin{aligned} f_2 &= 2f_1 \\ &= 2 \times 762 \times 10^{14} \text{ Hz} \\ &= 1524 \times 10^{14} \text{ Hz} \end{aligned}$$

But this frequency is outside the visible range, so we can't see it and chances are it won't go through the glass mirrors. So lasers usually only produce one frequency of light. But gas lasers can be built with special mirrors that allow many harmonics to be produced at once (e.g. CO₂ lasers).

The laser has an additional complication, and that is that it amplifies the light with a laser medium. That medium gives a new photon for every photon that passes through it, doubling the amount of light each time the wave passes through this *gain medium*. How that works is a subject for PH279. But for us, we can see that we can make standing waves in light.

Question 223.7.3

Question 223.7.4

Standing Waves in Rods and Membranes

singing Rod Demo

We have hinted all chapter that the analysis techniques we were building apply to structures. We need more math and computational tools to analyze complex structures

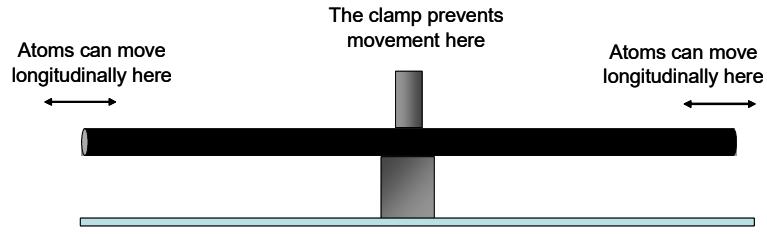
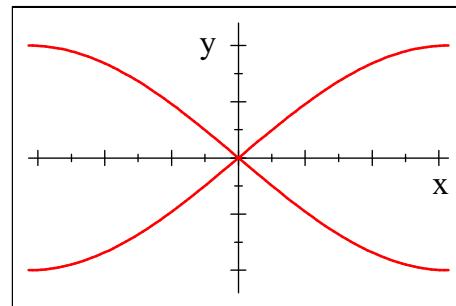


Figure 7.7.

like bridges and buildings, but we can tackle a simple structure like a rod that is clamped. The atoms in the rod can vibrate longitudinally. Since we have motion possible on both ends and not in the middle, we surmise that this system will have similar solutions as did the open ended pipe.

$$f_n = n \frac{v}{2L}$$

The fundamental looks like



But suppose we move the clamp. The clamp forces a node where it is placed. If we place the clamp at $L/4$

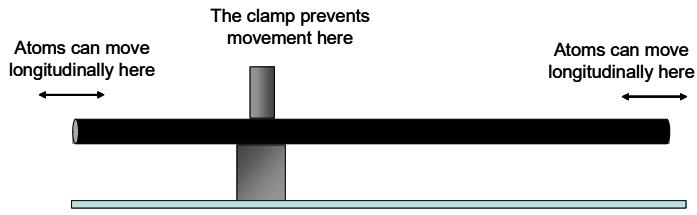
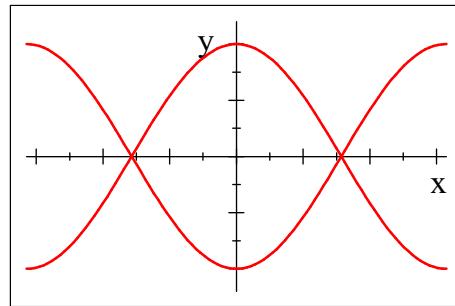


Figure 7.8.



We can perform a similar analysis for a drum head, but it is much more complicated. The modes are not points, but lines or curves, and the frequencies of oscillation are not integer multiples of each other. See for example <http://physics.usask.ca/~hirose/ep225/animation/drum/anim-drum.htm>.

Of course structures can also waggle on the ends. the ends can rotate counter to each other, etc. These are more complex modes than the longitudinal modes we have considered.

Question 223.7.5

8 Single Frequency Interference, Multiple Dimensions

Two speaker demo

So far we have had only waves mixed in a one-dimensional medium and we have only allowed for reflections to mix the waves. But surely we can make waves with different sources and mix them. Consider setting up two speakers playing the same frequency. We expect that we will still get regions of constructive and destructive interference. Where these regions will be really depends on the total phase difference between the two waves.

$$\begin{aligned}\Delta\phi &= (kx_2 - \omega t + \phi_2) - (kx_1 - \omega t + \phi_1) \\ &= k(x_2 - x_1) + (\phi_2 - \phi_1) \\ &= \frac{2\pi}{\lambda}(\Delta x) + \Delta\phi_o\end{aligned}$$

where we can see that there are at least two sources of phase difference here. One can be from the two waves traveling different paths and then combining (Δx) and the other is from them starting with a different phase to begin with $\Delta\phi$.

If we have two waves

$$\begin{aligned}y_1 &= A \sin(kx_1 - \omega t + \phi_1) \\ y_2 &= A \sin(kx_2 - \omega t + \phi_2)\end{aligned}$$

and we look at a particular part of the medium, that part will oscillate with an amplitude that depends on the relative starting points of the two waves, $\Delta\phi_o$ and on how the relative distances the waves have traveled to get to our particular location in the medium, Δx .

Fundamental Concepts

- In two dimensional problems, the total phase difference is given by $\Delta\phi = (2\pi\frac{\Delta r}{\lambda} + \Delta\phi_o)$
- In the total phase difference $\Delta\phi = \frac{2\pi}{\lambda}(\Delta x) + \Delta\phi_o$, the first term is due to path differences, the second to initial phase differences (whether the two mixed waves start together).

Mathematical treatment of single frequency interference

Question 223.8.1

Question 223.8.2

It is time to put our treatment of interference on a more general mathematical footing.

We start with two waves in the same medium

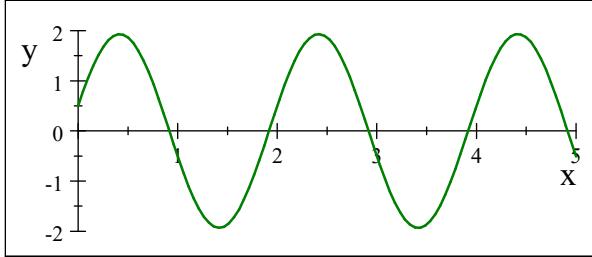
$$y_1 = y_{\max} \sin(kx_1 - \omega t + \phi_1)$$

$$y_2 = y_{\max} \sin(kx_2 - \omega t + \phi_2)$$

Each wave has its own phase constant. Each wave starts from a different position (one at x_1 and the other at x_2). The superposition yields.

$$y_r = y_{\max} \sin(kx_1 - \omega t + \phi_1) + y_{\max} \sin(kx_2 - \omega t + \phi_2)$$

which is graphed in the next figure.



Notice that the wave form is taller (larger amplitude). Noticed it is shifted along the x axis. This graph is not surprising to us now, because we have done a case like this before. We can find the shift in general rewriting y_r . We need a trig identity

$$\sin a + \sin b = 2 \cos\left(\frac{a-b}{2}\right) \sin\left(\frac{a+b}{2}\right)$$

then let $a = kx - \omega t$ and $b = kx - \omega t + \phi$

$$\begin{aligned}
 y_r &= y_{\max} \sin(kx_2 - \omega t + \phi_2) + y_{\max} \sin(kx_1 - \omega t + \phi_1) \\
 &= 2y_{\max} \cos\left(\frac{(kx_2 - \omega t + \phi_2) - (kx_1 - \omega t + \phi_1)}{2}\right) \sin\left(\frac{(kx_2 - \omega t + \phi_2) + (kx_1 - \omega t + \phi_1)}{2}\right) \\
 &= 2y_{\max} \cos\left(\frac{kx_2 - kx_1}{2} + \frac{\phi_2 - \phi_1}{2}\right) \sin\left(\frac{kx_2 + kx_1 - 2\omega t + \phi_2 + \phi_1}{2}\right) \\
 &= 2y_{\max} \cos\left(k\frac{x_2 - x_1}{2} + \frac{\phi_2 - \phi_1}{2}\right) \sin\left(k\frac{x_2 + x_1}{2} - \omega t + \frac{\phi_2 + \phi_1}{2}\right) \\
 &= 2y_{\max} \cos\left(k\frac{\Delta x}{2} + \frac{\Delta\phi_o}{2}\right) \sin\left(k\frac{x_2 + x_1}{2} - \omega t + \frac{\phi_2 + \phi_1}{2}\right) \\
 &= 2y_{\max} \cos\left(\frac{1}{2} \left(\frac{2\pi}{\lambda} \Delta x + \Delta\phi_o \right)\right) \sin\left(k\frac{x_2 + x_1}{2} - \omega t + \frac{\phi_2 + \phi_1}{2}\right) \\
 &= 2y_{\max} \cos\left(\frac{1}{2} (\Delta\phi)\right) \sin\left(k\frac{x_2 + x_1}{2} - \omega t + \frac{\phi_2 + \phi_1}{2}\right)
 \end{aligned}$$

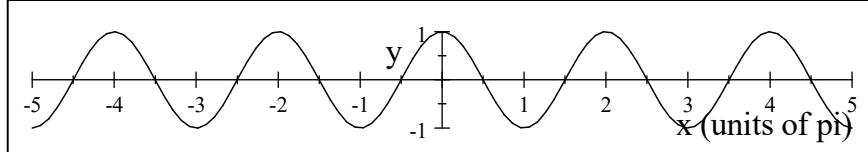
where the last line is just a rearrangement to match the form we got last time we did this problem with just one phase constant. Clearly, we see the amplitude depends on what we call the phase difference. And we can see our two sources of phase difference. One can be from the two waves traveling different paths and then combining (Δx) and the other is from the two waves starting with a different phase to begin with, $\Delta\phi_o$. If the total phase difference between the two waves is a multiple of 2π , then the two waves will experience constructive interference

$$\Delta\phi = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots$$

Let's see that this works. Our amplitude is

$$A = 2y_{\max} \cos\left(\frac{1}{2}(\Delta\phi)\right)$$

and if we look at a cosine function we see that $\cos(\theta)$ is either 1 or -1 at $\theta = n\pi$.



So if $\Delta\phi = m2\pi$ then the amplitude is

$$\begin{aligned} A &= 2y_{\max} \left(\frac{1}{2}(m2\pi) \right) \\ &= 2y_{\max} \cos(m\pi) \end{aligned}$$

We don't really care if the amplitude function is big positively or negatively. So we get constructive interference for either 1 or -1 . Then, this our case for constructive interference.

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta x + \Delta\phi_o \right) = n2\pi \quad n = 0, \pm 1, \pm 2, \pm 3, \dots$$

How about for destructive interference? We start again with our amplitude function

$$A = 2y_{\max} \cos\left(\frac{1}{2}(\Delta\phi)\right)$$

but now we want when the cosine part is zero.

$$\cos(\theta) = 0$$

Looking at our cosine graph again, that happens for cosine when $\theta = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \dots$. We could write this as $\theta = (m + \frac{1}{2})\pi$ for $m = 0, 1, 2, \dots$. But remember that in our amplitude function, we already have the $1/2$ in the function, so we want $\Delta\phi$ to have just the odd integer multiple of π . We could write this as

$$\Delta\phi = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots$$

So our condition for destructive interference is

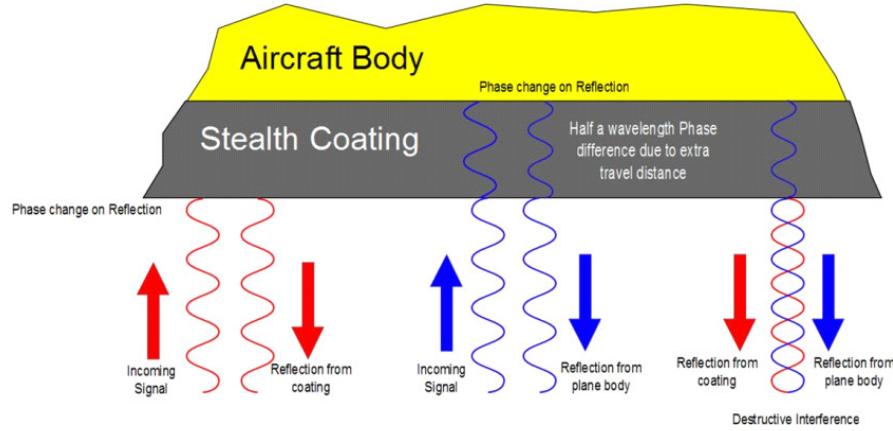
$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots$$

We have developed a useful matched set of equations that will tell us if we mix two waves when we will have constructive and destructive interference:

$$\begin{aligned}\Delta\phi &= \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots && \text{Constructive} \\ \Delta\phi &= \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots && \text{Destructive}\end{aligned}$$

Let's take an example to see how this can be used.

Example of two wave interference: Stealth Fighter



The stealth fighter is coated with an anti-reflective polymer. This is part of its mechanism for making the plane invisible to radar. Suppose we have a radar system with a wavelength of 3.00 cm. Further suppose that the index of refraction of the anti-reflective polymer is $n = 1.50$, and that the aircraft index of refraction is very large, how thick would you make the coating?

We want destructive interference, so let's start with our destructive interference condition

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad \text{Destructive}$$

The radar waves all hit the plane in phase. From the figure, we see that the radar wave will reflect off of the coating. Because the index of refraction of the coating is large, this is like a fixed end of a rope. There will be an inversion.

But some of the wave will penetrate the polymer. This will reflect off of the plane body. The plane body has a very large index of refraction, so once again the wave will experience an inversion. The outgoing waves would then be in phase and create constructive interference because

$$\Delta\phi_o = \pi - \pi = 0$$

at this point. Thus

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad \text{Destructive}$$

But we have to remember the path difference! The part of the wave that entered the polymer travels farther. If that path difference, Δr , is just right so that

$$\frac{2\pi}{\lambda} \Delta r = \pi$$

This is the $m = 0$ case in our destructive interference case

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r \right) = (2(0) + 1)\pi = \pi$$

then the amplitude function would be

$$\begin{aligned} A &= 2E_{\max} \cos \left(\frac{2\pi}{\lambda} \Delta r \right) \\ &= 2E_{\max} \cos \left(\frac{1}{2}(\pi) \right) \\ &= 0 \end{aligned}$$

and we have destructive interference. Note that these are electromagnetic waves, so instead of y_{\max} we have used E_{\max} as the individual wave amplitude. But the important thing is that the plane cannot be seen by the radar! Of course, this works for $m = 1$ and $m = 2$, etc. as well. Any odd multiple of π will work.

$$\frac{2\pi}{\lambda} \Delta r = (2m + 1)\pi$$

where $m = 0, 1, 2, \dots$ so that we are guaranteed an odd multiple of π . This is our condition for destructive interference.

But we are interested in the thickness. We realize that Δr is about twice the thickness, since the wave travels through the coating and back. So let's let $\Delta r \approx 2t$

$$\begin{aligned} \frac{2\pi}{\lambda} 2t &\approx (2m + 1)\pi \\ 2t &\approx (2m + 1) \frac{\lambda}{2} \\ 2t &\approx \left(m + \frac{1}{2} \right) \lambda \end{aligned}$$

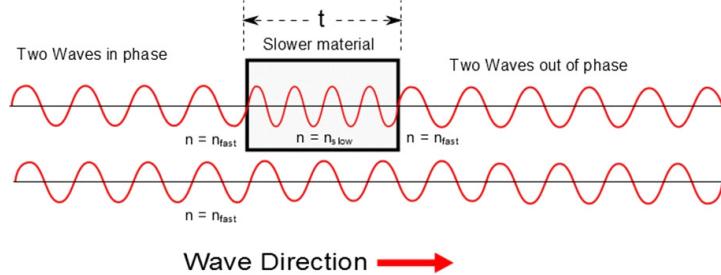
$$t \approx \left(m + \frac{1}{2} \right) \frac{\lambda}{2}$$

But there is a further complication. We should write our thickness equation as

$$t \approx \left(m + \frac{1}{2} \right) \frac{\lambda_{in}}{2}$$

because Δr has to provide an odd integer times the wavelength *inside the coating* for the phase to be right. After all, the wave is traveling inside the coating. We know that the wavelength will change as we enter the slower material.

To see this, consider two waves traveling to the right. One passes through a slower medium. We expect the wavelength to shorten. We can see that, depending on the thickness t , the wave may be in phase or out of phase. In the next picture, the thickness is just right so that we have destructive interference.



We have such a wavelength shift in the coating. But we don't know the wavelength inside the coating. All we know is the radar wavelength, λ_{out} . We can fix it by writing the wavelength inside in terms of the wavelength outside. Earlier in our studies we found that the new wavelength will be given by equation (4.3)

$$\lambda_f = \frac{v_f}{v_i} \lambda_i$$

Let's rewrite this for our case

$$\lambda_{in} = \frac{v_{in}}{v_{out}} \lambda_{out}$$

We can express this in terms of the index of refraction

$$n = \frac{c}{v}$$

by multiplying the left hand side by c/c then

$$\lambda_{in} = \frac{cv_{in}}{cv_{out}} \lambda_{out}$$

or

$$\begin{aligned} \lambda_{in} &= \frac{\frac{c}{v_{out}}}{\frac{c}{v_{in}}} \lambda_{out} \\ &= \frac{n_{out}}{n_{in}} \lambda_{out} \end{aligned}$$

in the case of our aircraft coating the outside medium is air so $n_{out} \approx 1$

$$\lambda_{in} = \frac{1}{n_{in}} \lambda_{out}$$

This is this wavelength we need to match as the radar signal enters the medium.

Using this expression for λ_{in} in

$$t \approx \left(m + \frac{1}{2} \right) \frac{\lambda_{in}}{2} \quad m = 0, 1, 2, \dots$$

will give us the condition for destructive interference. Let's rewrite our λ_{in} equation for our case of a coating and air

$$\lambda_{in} = \lambda_{coating} = \frac{1}{n_{in}} \lambda_{out} = \frac{1}{n_{coating}} \lambda_{air}$$

thus

$$t \approx \left(m + \frac{1}{2} \right) \frac{1}{2} \left(\frac{\lambda_{air}}{n_{coating}} \right) \quad m = 0, 1, 2, \dots$$

is our condition for being stealthy.

Let's assume we want the thinnest coating possible, so we set $m = 0$. Then

$$t \approx \left(\frac{1}{4} \right) \left(\frac{\lambda_{air}}{n_{coating}} \right)$$

and our thickness would be

$$t \approx \left(\frac{1}{4} \right) \left(\frac{3.00 \text{ cm}}{1.50} \right) = 0.5 \text{ cm}$$

This seems doable for an aircraft coating!

Of course we could also make a plane that would be more visible to radar by choosing the constructive interference case. Suppose we are building a search and rescue plane. We want to enhance it's ability to be seen by radar in fog. We start with the condition for constructive interference

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad \text{Constructive}$$

It will still be true that $\Delta\phi_o = 0$.

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r \right) = m2\pi$$

and it is still true that $\Delta r \approx 2t$.

$$\left(\frac{2\pi}{\lambda} 2t \right) \approx m2\pi$$

then

$$\begin{aligned} \left(\frac{1}{\lambda} 2t \right) &\approx m \\ t &\approx \frac{1}{2} m \lambda \end{aligned}$$

and we still have to adjust for the coating index of refraction

$$t \approx \frac{m}{2} \left(\frac{\lambda_{air}}{n_{coating}} \right)$$

And once again we have several choices for m

$$t \approx \frac{m}{2} \left(\frac{\lambda_{air}}{n_{coating}} \right) \quad m = 0, 1, 2, \dots$$

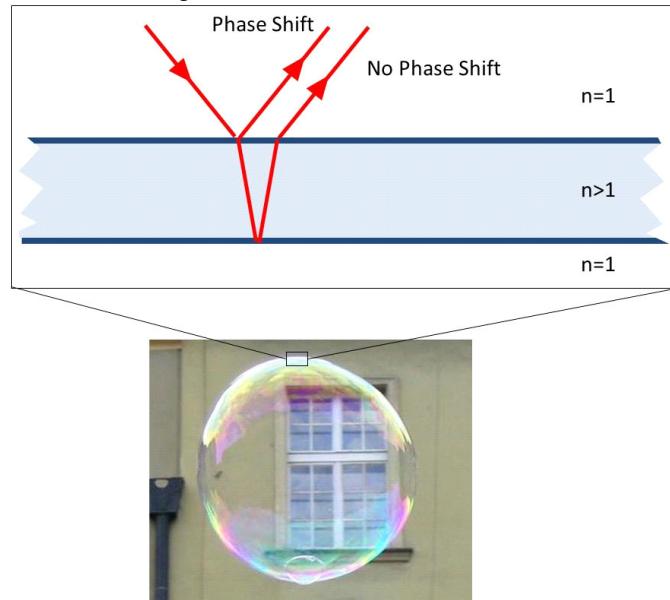
But now the coating will provide constructive interference, making it easier to track on radar from the command center. For the thinnest possibility, set $m = 1$ because the $m = 0$ case doesn't give us any thickness.

$$\begin{aligned} t &\approx m \frac{1}{2} \left(\frac{\lambda_{air}}{n_{coating}} \right) \\ &= \frac{1}{2} \left(\frac{3.00 \text{ cm}}{1.50} \right) \\ &= 1 \text{ cm} \end{aligned}$$

Note that we reasoned out these equations for the boundary conditions that we have in our problem. If the boundary conditions change, so do the equations.

Example of two wave interference: soap bubble

Take a soap bubble for example.



Interference from a soap bubble. (Bubble image in the Public Domain, courtesy Marcin Deregowski)

Now we have a phase shift on the first reflection, but not one on the reflection from the inside surface of the bubble because the bubble is full of air. The index of refraction

of air is less than that for the bubble material. So as we leave the bubble material it is more like having a free end of a rope. As the waves leave the surface, they are half a wavelength out of phase due to $\Delta\phi_o$ because of the single inversion from the bubble outer surface. We would have destructive interference due to just this, but we also have to account for the bubble thickness. If this thickness is a multiple of a wavelength, then we are still have half a wavelength out of phase and we have destructive interference.

Here are our basic equations

$$\begin{aligned}\Delta\phi &= \left(\frac{2\pi}{\lambda}\Delta r + \Delta\phi_o\right) = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \text{ Constructive} \\ \Delta\phi &= \left(\frac{2\pi}{\lambda}\Delta r + \Delta\phi_o\right) = (2m+1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \text{ Destructive}\end{aligned}$$

Suppose we want want constructive interference to get our colors, so we take the first

$$\Delta\phi = \left(\frac{2\pi}{\lambda}\Delta r + \Delta\phi_o\right) = m2\pi$$

and this time we have

$$\begin{aligned}\Delta\phi_o &= \phi_{transmitted} - \phi_{reflected} \\ &= 0 - \pi \\ &= -\pi\end{aligned}$$

It is still true that $\Delta r \approx 2t$ so from our constructive interference equation

$$\begin{aligned}\frac{2\pi}{\lambda}2t - \pi &= m2\pi \\ \frac{2}{\lambda}t - \frac{1}{2} &= m \\ \frac{2}{\lambda}t &= m + \frac{1}{2} \\ t &= \frac{\lambda}{2} \left(m + \frac{1}{2}\right)\end{aligned}$$

We again have the problem that this wavelength must be the wavelength inside the bubble material $\lambda = \lambda_{in}$. But we see the outside wavelength λ_{out} . We can reuse our conversion from outside to inside wavelength from our last problem because we are once again in air and $n_{air} \approx 1$.

$$\lambda_{in} = \frac{1}{n_{in}}\lambda_{out}$$

then

$$t = \frac{\lambda_{out}}{2n_{in}} \left(m + \frac{1}{2}\right) \quad m = 0, 1, 2, \dots$$

Or writing this with $n_{in} = n_{bubble}$ to make it clear that the inside material is the bubble solution,

$$t = \left(m + \frac{1}{2}\right) \frac{1}{2} \left(\frac{\lambda_{air}}{n_{bubble}}\right) \quad m = 0, 1, 2, \dots$$

but this was the equation for destructive interference for the plane! We can see that memorizing the thickness equations won't work. We need to start with our conditions on $\Delta\phi$ for constructive and destructive interference to be safe!

How about the dark parts of the bubble with no color (the parts we can see through). These would be destructive interference

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta x + \Delta\phi_o \right) = (2m + 1)\pi$$

We can fill in the pieces to obtain

$$\begin{aligned} \left(\frac{2\pi}{\lambda} 2t - \pi \right) &= (2m + 1)\pi \\ \frac{2}{\lambda} 2t - 1 &= (2m + 1) \\ \frac{2}{\lambda} 2t &= (2m + 1) + 1 \\ \frac{4}{\lambda} t &= (2m + 1) + 1 \\ t &= \frac{\lambda}{4} (2m + 2) \\ t &= \frac{\lambda}{2} (m + 1) \\ t &= \frac{m + 1}{2} \left(\frac{\lambda_{out}}{n_{bubble}} \right) \end{aligned}$$

This is our condition for destructive interference for the bubble. We don't have to, but we could write $m + 1 = p$ where p is an integer that starts at 1 instead of zero.

$$t = \frac{p}{2} \left(\frac{\lambda_{out}}{n_{bubble}} \right) \quad p = 1, 2, \dots$$

But this is very like the condition for constructive interference for the plane.

Hopefully, it is apparent that we have to start with our basic equations

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta x + \Delta\phi_o \right) = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \text{ Constructive}$$

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta x + \Delta\phi_o \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \text{ Destructive}$$

each time we attempt an interference problem because the outcome depends on both Δx and $\Delta\phi_o$. We have to construct the equation each time for the interference condition we want (constructive or destructive) finding Δx and $\Delta\phi_o$ for the boundary conditions we have.

Single frequency interference in more than one dimension

Two speaker demo

Suppose I put two speakers facing each other 3 m apart. And suppose I want four nodes in the middle so we can easily find them. Then I will need

$$2\lambda = 3 \text{ m}$$

or $\lambda = \frac{3}{2} \text{ m}$. If we are at about 20°C then $v = 343 \text{ m/s}$ and

$$\begin{aligned} f &= \frac{343 \text{ m/s}}{\frac{3}{2} \text{ m.}} \\ &= 228.67 \text{ Hz} \end{aligned}$$

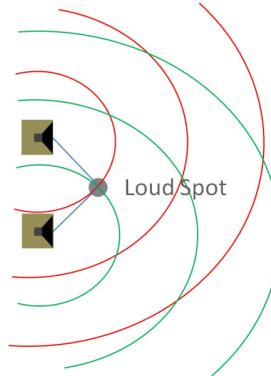
Are the nodes symmetrically placed?

Question 223.8.3

Question 223.8.4

If $\Delta\phi_o = 0$ the nodes should be spaced symmetrically between the two speakers. The rest of the phase comes from the difference in starting positions.

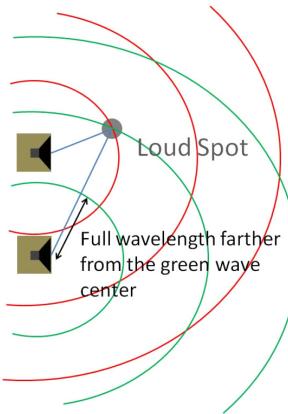
But what happens if our waves don't travel along the same line? Suppose you are at a dance, and there are two speakers. Further suppose that you are testing the system with a constant tone (either that, or you have somewhat boring music with constant tones). Suppose the two speakers make waves in phase. If you are equal distance from the two speakers, you would expect constructive interference because both $\Delta\phi_o = 0$ and $\Delta x = 0$ for this case.



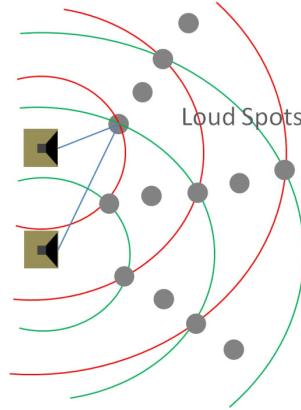
But there are more places where we expect constructive interference, because we know the sound wave is really spherical. Any time the path difference, $\Delta x = n\lambda$, then

$$\Delta\phi = \frac{2\pi}{\lambda} (n\lambda) = n2\pi$$

and we will have constructive interference. The next figure shows an example where the path difference is one wavelength.



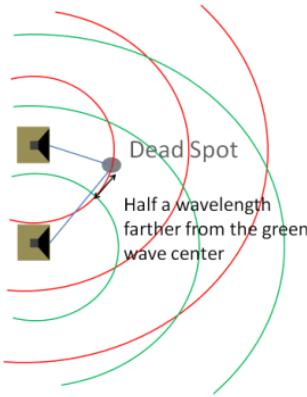
But any of these spots will experience constructive interference. Note the loud spots are where there are two crests or two troughs together.



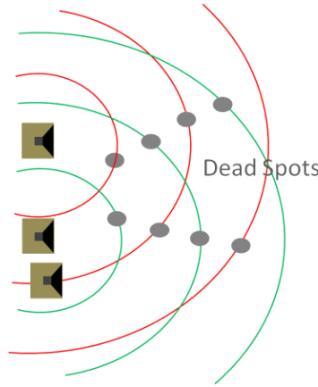
We also expect to see destructive interference. This should occur where path differences are multiples of $\Delta x = \lambda/2$ so that

$$\Delta\phi = \frac{2\pi}{\lambda} \left(n \frac{\lambda}{2} \right) = n\pi$$

The situation of being just half a wavelength off is shown next



but there are many places where we could be a multiple of a wavelength plus and extra half a wavelength off. Each of these will produce destructive interference.



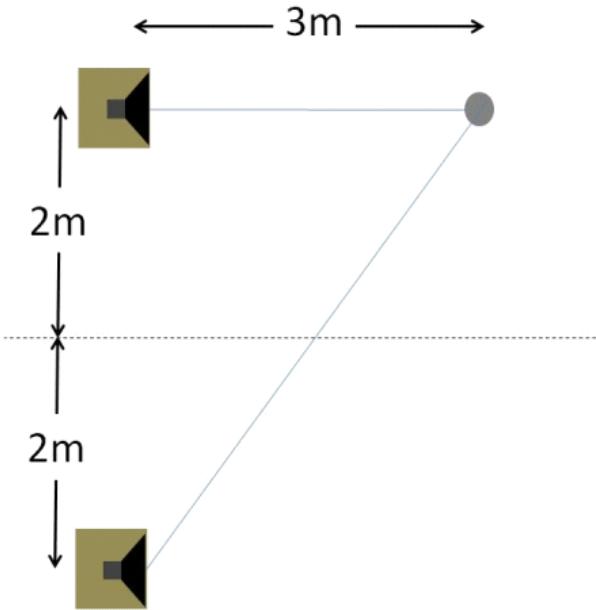
Recall that when you moved from one dimension to two dimensions in PH 121 or Dynamics problems, you changed from the variables x and y to the variable r where

$$r = \sqrt{x^2 + y^2}$$

Thus our phase becomes

$$\Delta\phi = \left(2\pi \frac{\Delta r}{\lambda} + \Delta\phi_o\right)$$

In our dance example, suppose we have speakers that are 4 m apart and we are standing 3 m directly in front of one of the speakers. Further suppose that we play an *A* just above middle *C* which has a frequency of 440 Hz. The speed of sound is 343 m/s. Our speakers are connected to the same stereo with equal length wires. What is the phase difference at this spot?



From the geometry we can tell that the path from the second speaker must be 5 m. So

$$\begin{aligned}\Delta r &= 5 \text{ m} - 3 \text{ m} \\ &= 2 \text{ m}\end{aligned}$$

We can tell that the wavelength is

$$\begin{aligned}\lambda &= \frac{v}{f} \\ &= \frac{343 \text{ m/s}}{440 \text{ Hz}} \\ &= 0.77955 \text{ m}:\end{aligned}$$

Since the speakers are connected to the same stereo with equal length wires, $\Delta\phi_o = 0$.

Then

$$\begin{aligned}\Delta\phi &= \frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \\ &= \frac{2\pi}{0.77955 \text{ m}} (2 \text{ m}) + 0 \\ &= 5.1312\pi \\ &= 2\pi + 3.1312\pi\end{aligned}$$

We should ask, is this constructive or destructive interference? Well, it is neither purely constructive interference nor total destructive interference. Our amplitude would be

$$2A \cos\left(\frac{1}{2} \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right)\right)$$

so in this case we get

$$2A \cos\left(\frac{1}{2}(2\pi + 3.1312\pi)\right) = -0.40927A$$

which is smaller (in magnitude) than A , so it is partial destructive interference. It would be quieter at this spot than if we had just one speaker operating.

You might guess that this sort of analysis plays a large part in design of concert halls. It also is important in mechanical designs.

But you should have seen a deficit in what we have learned so far. Up to this point, we have only mixed waves that have the same frequency. Can we mix waves that have different frequencies? That will be the subject of our next lecture.

9 Multiple Frequency Interference

Fundamental Concepts

- Mixing waves of different frequencies produces a time-varying amplitude called beating.
- Complex waves can be treated as a superposition of simple sinusoidal waves.
- Limited signals are multi-frequency

Beats

Question 223.9.1

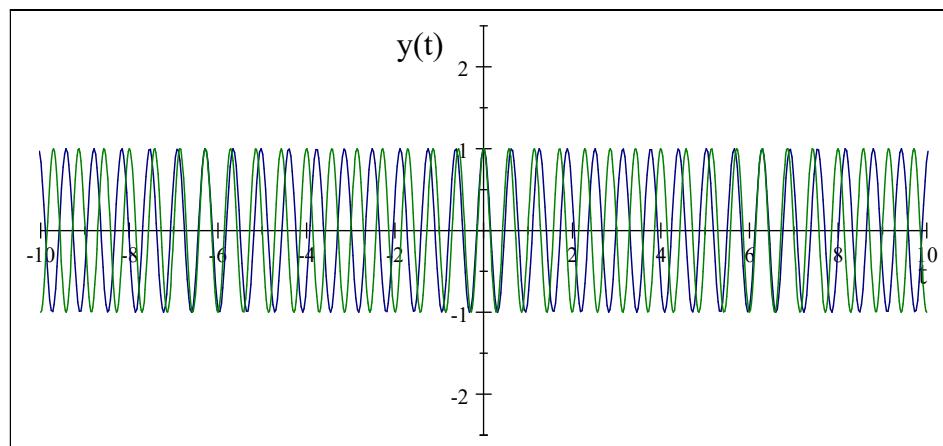
Beat Demo

Up till now we only superposed waves that had the same frequency. But what happens if we take waves with different frequencies?

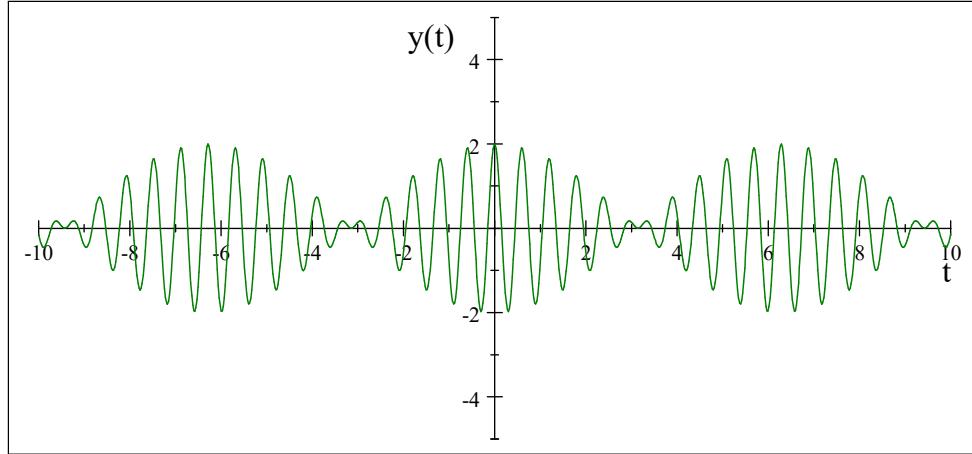
$$y_1 = y_{\max} \sin(kx - \omega_1 t)$$

$$y_2 = y_{\max} \sin(kx - \omega_2 t)$$

We can plot both waves on the same graph, in this case a history graph.



Notice that there are places where the waves are in phase, and places where they are not. The superposition looks like this



where there is constructive interference, the resulting wave amplitude is large, where there is destructive interference, the resulting amplitude is zero. We get a traveling wave who's amplitude varies. We can find the amplitude function algebraically.

We can write these as

$$y_1 = y_{\max} \sin(kx - 2\pi f_1 t)$$

$$y_2 = y_{\max} \sin(kx - 2\pi f_2 t)$$

The sum is just

$$y = y_{\max} \sin(kx - 2\pi f_1 t) + y_{\max} \sin(kx - 2\pi f_2 t)$$

We use another trig identity

$$\sin(a) + \sin(b) = 2 \cos\left(\frac{a-b}{2}\right) \sin\left(\frac{a+b}{2}\right)$$

which allows us to write this as

$$\begin{aligned} y &= 2y_{\max} \cos\left(\frac{kx - 2\pi f_2 t - (kx - 2\pi f_1 t)}{2}\right) \sin\left(\frac{kx - 2\pi f_2 t + kx - 2\pi f_1 t}{2}\right) \\ &= 2y_{\max} \cos\left(2\pi \frac{f_1 - f_2}{2} t\right) \sin\left(-2\pi \frac{f_1 + f_2}{2} t\right) \\ &= \left[2y_{\max} \cos\left(2\pi \frac{f_1 - f_2}{2} t\right)\right] \sin\left(kx - 2\pi \frac{f_1 + f_2}{2} t\right) \end{aligned}$$

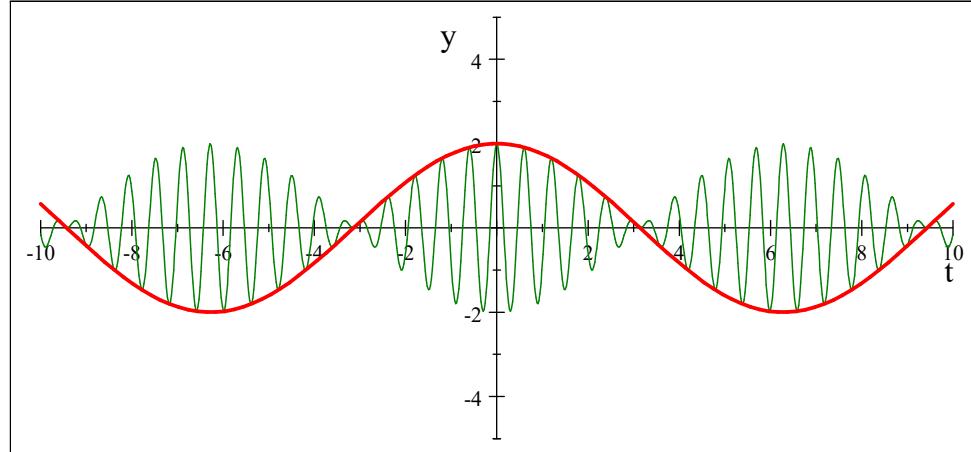
We see that we have a part that has a frequency that is the average of f_1 and f_2 . This is the frequency we hear. But we have another complicated amplitude term, and this time it is a function of time (just to be confusing). The amplitude has its own frequency that

is half the difference of f_1 and f_2 .

$$A_{\text{resultant}} = 2y_{\max} \cos \left(2\pi \frac{f_1 - f_2}{2} t \right)$$

So the sound amplitude will vary in time for a given spatial location.

The situation is odder still. We have a cosine function, but it is really an envelope for the higher frequency motion of the air particles.



Our ear drum does not care which way the envelope function goes. We can see that the green (thin line) wave will push and pull air molecules, and therefore our ear drums, with maximum loudness at twice this frequency. So we will hear two maxima for every envelope period!

This frequency with which we hear the sound get loud at a given location as the wave goes by is called the *beat frequency*. The red envelope (solid heavy line in the last figure) has a frequency of

$$f_A = \frac{f_1 - f_2}{2}$$

So our beat frequency is

$$f_{\text{beat}} = |f_1 - f_2|$$

Question 223.9.2

Non Sinusoidal Waves

You have probably wondered if all waves are sinusoidal. Can the universe really be described by such simple mathematics? The answer is both no, and yes. There are non-sinusoidal waves, in fact, most waves are not sinusoidal. But it turns out that we can use a very clever mathematical trick to make any shape wave out of a superposition

of many sinusoidal waves. So our mathematics for sinusoidal waves turns out to be quite general.

Music and Non-sinusoidal waves

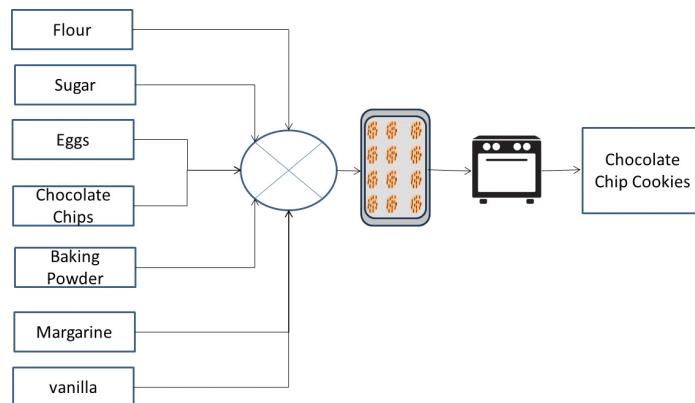
Let's take the example of music.

From our example of standing waves on strings, we know that a string can support a series of standing waves with discrete frequencies—the harmonic series. We have also discussed that usually we excite the waves with a pluck or some discrete event, not with an oscillator. Only the harmonic series of frequencies will resonate, creating standing waves. Other frequency waves die out quickly. But there is no reason to suppose that we get energy in only one standing wave at a time. Most sounds are a combination of harmonics.

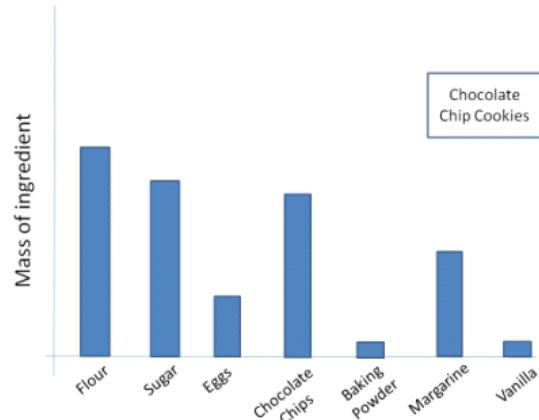
The fundamental mode tends to give us the pitch we hear, but what are the other standing waves for?

To understand, let's take an analogy. Making cookies and cakes.

Here is the beginning of a recipe for cookies.

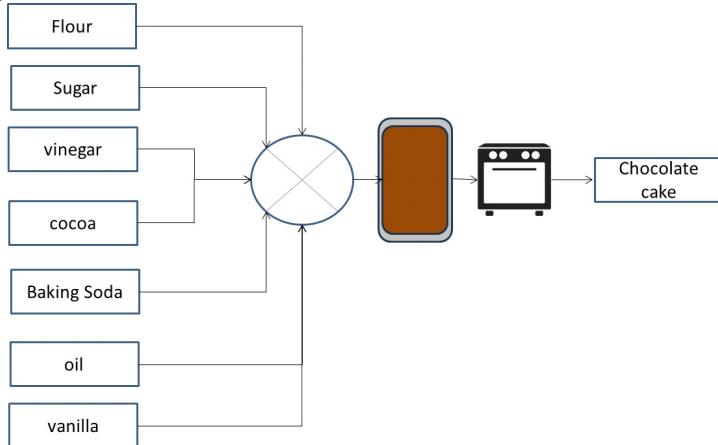


The recipe is a list of ingredients, and a symbolic instruction to mix and bake. The product is chocolate chip cookies. Of course we need more information. We need to know how much of each ingredient to use.



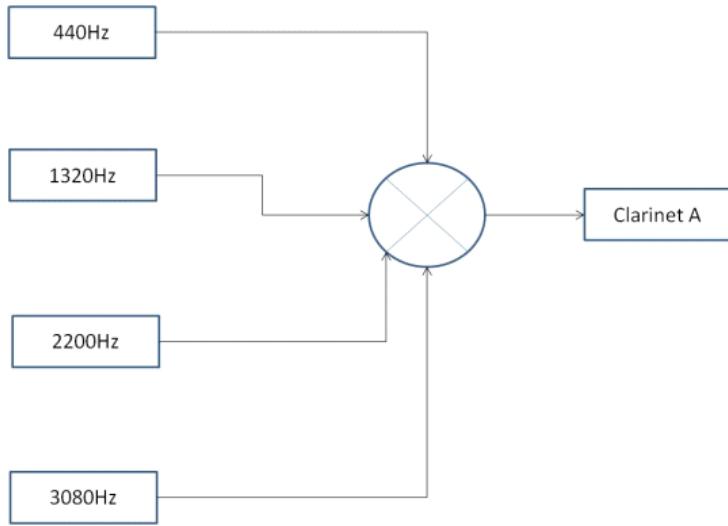
This graph gives us the amount of each ingredient by mass.

Now suppose we want chocolate cake.

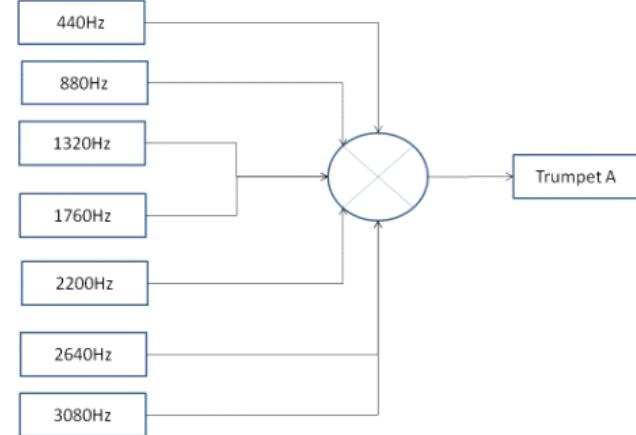


The predominant taste in each of these foods is chocolate. But chocolate cake and chocolate chip cookies don't taste exactly the same. We can easily see that the differences in the other ingredients make the difference between the "cookie" taste and the "cake" taste that goes along with the "chocolate" taste that predominates.

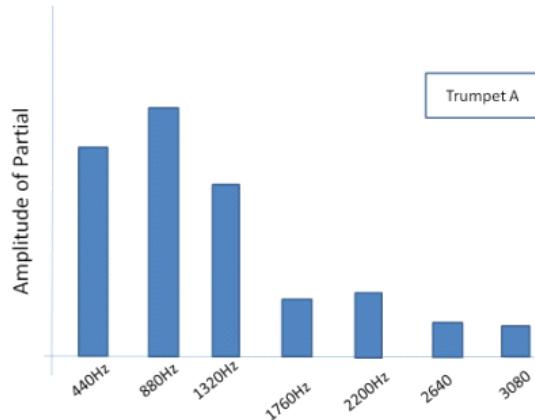
The sound waves produced by musical instruments work in a similar way. Here is a recipe for an "A" note from a clarinet.



and here is one for a trumpet playing the same “A” note.

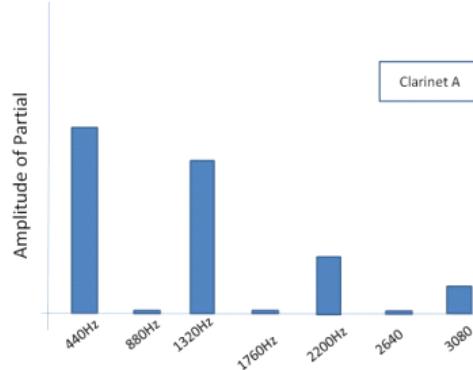


A trumpet sounds different than a clarinet, and now we see why. There are more harmonics involved with the trumpet sound than the clarinet sound. These extra standing waves make up the “brassiness” of the trumpet sound. As with our baking example, we need to know how much of each standing wave we have. Each will have a different amplitude. For our trumpet, we might get amplitudes as shown.



Note that the second harmonic has a larger amplitude, but we still hear the musical note as “A” at 440 Hz. A Flugelhorn horn would still sound brassy, but would have a different mix of harmonics.

The clarinet graph would look quite different, perhaps something like this



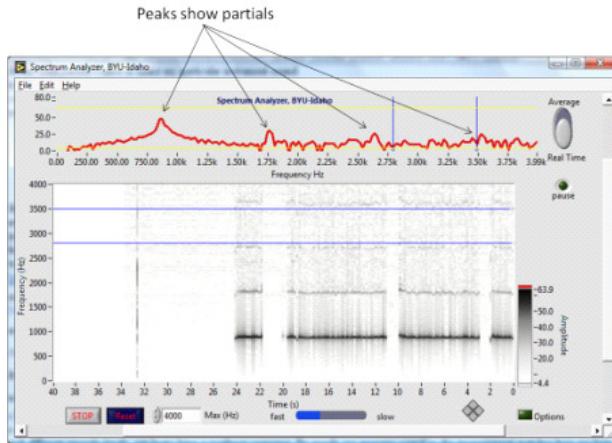
because it does not have as many “ingredients” as the trumpet.

All of this should remind you of our analysis of open and closed pipes. Remember when we closed a pipe, we lost all the even multiples the fundamental frequency. A similar thing is happening with our instruments. The rich sound of the brass instrument includes more harmonics and this is achieved by the shape of the instrument (the flared bell is a big part of making these extra harmonics and providing the rich trumpet sound).

Spectrometer Demo

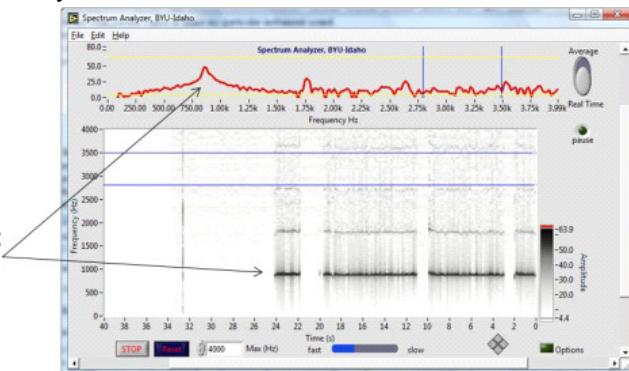
We have a tool that you can download to your PC to detect the mix of harmonics of musical instruments, or mechanical systems. In music, the different harmonics are

called *partials* because they make up part of the sound. A graph that shows which harmonics are involved is called a *spectrum*. The next figure is the spectrum of a six holed bamboo flute. Note that there are several harmonics involved.

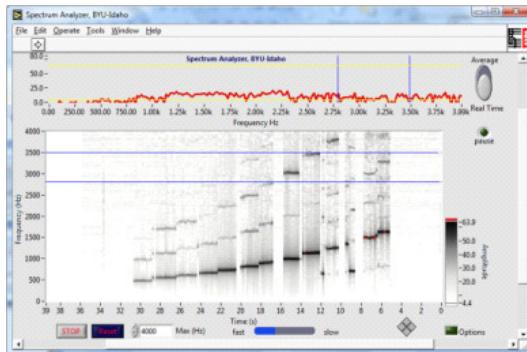


Note that our software display has two parts. One is the instantaneous spectrum, and one is the spectrum time history.

Notice that there
are non-harmonic
partials in this
instrument too!

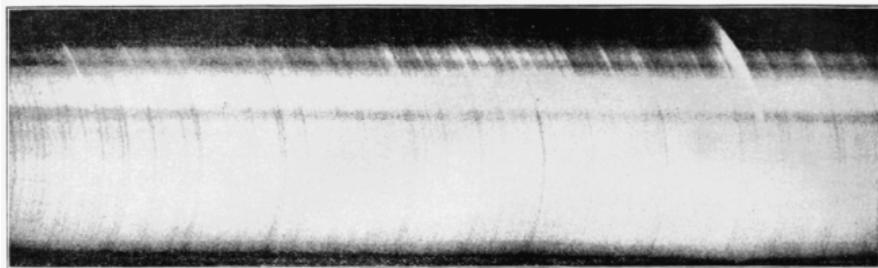


By observing the time history, we can see changes in the spectrum. We can also see that we don't have pure harmonics. The graph shows some response off the specific harmonic frequencies. This six holed flute is very "breathy" giving a lot of wind noise along with the notes, and we see this in the spectrum. In the next picture, I played a scale on the flute.



The instantaneous spectrum is not active in this figure (since it can't show more than one note at a time on the instantaneous graph) but in the time history we see that as the fundamental frequency changes by shorting the length of the flute (uncovering holes), we see that every partial also goes up in frequency. The flute still has the characteristic spectrum of a flute, but shifted to new set of frequencies. We can use this fact to identify things by their vibration spectrum. In fact, that is how you recognize voices and instruments within your auditory system!

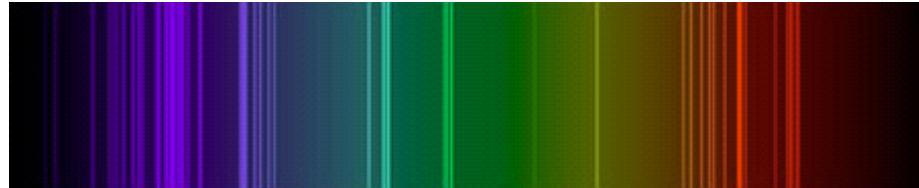
The technique of taking apart a wave into its components is very powerful. With light waves, the spectrum is an indication of the chemical composition of the emitter. For example, the spectrum of the sun looks something like this



Solar coronal spectrum taken during a solar eclipse. The successive curved lines are each different wavelengths, and the dark lines are wavelengths that are absorbed. The pattern of absorbed wavelengths allows a chemical analysis of the corona. (Image in the Public Domain, originally published in Bailey, Solon, L, *Popular Science Monthly*, Vol 60, Nov. 1919, pp 244)

The lines in this graph show the amplitude of each harmonic component of the light. Darker lines have larger amplitudes. The harmonics come from the excitation of electrons in their orbitals. Each orbital is a different energy state, and when the electrons jump from orbital to orbital, they produce specific wave frequencies. By observing the mix of dark lines in previous figure, and comparing to laboratory measurements from

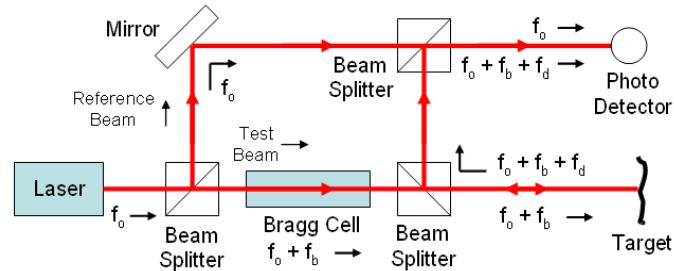
each element (see next figure) we can find the composition of the source. This figure shows the emission spectrum for Calcium. Because it is an emission spectrum the lines are bright instead of dark. We can even see the color of each line!



Emission spectrum of Calcium (Image in the Public Domain, courtesy NASA)

Vibrometry

Just like each atom has a specific spectrum, and each instrument, each engine, machine, or anything that vibrates has a spectrum. We can use this to monitor the health of machinery, or even to identify a piece of equipment. Laser or acoustic vibrometers are available commercially.



Laser Vibrometer Schematic (Public Domain Image from Laderaranch:
http://commons.wikimedia.org/wiki/File:LDV_Schematic.png)

They provide a way to monitor equipment in places where it would be dangerous or even impossible to send a person. The equipment also does not need to be shut down, a great benefit for factories that are never shut down, or for a satellite system that cannot be reached by anyone.

Fourier Series: Mathematics of Non-Sinusoidal Waves

We should take a quick look at the mathematics of non-sinusoidal waves.

Let's start with a superposition of many sinusoidal waves. The math looks like this

$$y(t) = \sum_n (A_n \sin(2\pi f_n t) + B_n \cos(2\pi f_n t))$$

where A_n and B_n are a series of coefficients and f_n are the harmonic series of frequencies. The coefficients are amplitudes for the many individual waves making up the complicated wave.

Example: Fourier representation of a square wave.

For example, we could represent a function $f(x)$ with the following series

$$f(x) = C_0 + C_1 \cos\left(\frac{2\pi}{\lambda}x + \varepsilon_1\right) \quad (9.1)$$

$$+ C_2 \cos\left(\frac{2\pi}{\frac{\lambda}{2}}x + \varepsilon_2\right) \quad (9.2)$$

$$+ C_3 \cos\left(\frac{2\pi}{\frac{\lambda}{3}}x + \varepsilon_3\right) \quad (9.3)$$

$$+ \dots \quad (9.4)$$

$$+ C_n \cos\left(\frac{2\pi}{\frac{\lambda}{n}}x + \varepsilon_n\right) \quad (9.5)$$

$$+ \dots \quad (9.6)$$

where we will let $\varepsilon_i = \omega_i t + \phi_i$

The C 's are just coefficients that tell us the amplitude of the individual cosine waves. The more terms in the series we take, the better the approximation we will have, with the series exactly matching $f(x)$ when the number of terms, $N \rightarrow \infty$.

Usually we rewrite the terms of the series as

$$C_m \cos(mkx + \varepsilon_m) = A_m \cos(mkx) + B_m \sin(mkx) \quad (9.7)$$

where k is the wavenumber

$$k = \frac{2\pi}{\lambda} \quad (9.8)$$

and λ is the wavelength of the complicated but still periodic function $f(x)$. Then we identify

$$A_m = C_m \cos(\varepsilon_m) \quad (9.9)$$

$$B_m = -C_m \sin(\varepsilon_m) \quad (9.10)$$

then

$$f(x) = \frac{A_o}{2} + \sum_{m=1}^{\infty} A_m \cos(mkx) + \sum_{m=1}^{\infty} B_m \sin(mkx) \quad (9.11)$$

where we separated out the $A_o/2$ term because it makes things nicer later.

Fourier Analysis

The process of finding the coefficients of the series is called *Fourier analysis*. We start by integrating equation (9.11)

$$\int_0^\lambda f(x) dx = \int_0^\lambda \frac{A_o}{2} dx + \int_0^\lambda \sum_{m=1}^{\infty} A_m \cos(mkx) dx + \int_0^\lambda \sum_{m=1}^{\infty} B_m \sin(mkx) dx \quad (9.12)$$

We can see immediately that all the sine and cosine terms integrate to zero (we integrated over a wavelength) so

$$\int_0^\lambda f(x) dx = \int_0^\lambda \frac{A_o}{2} dx = \frac{A_o}{2} \lambda \quad (9.13)$$

We solve this for A_o

$$A_o = \frac{2}{\lambda} \int_0^\lambda f(x) dx \quad (9.14)$$

To find the rest of the coefficients we need to remind ourselves of the orthogonality of sinusoidal functions

$$\int_0^\lambda \sin(akx) \cos(bkx) dx = 0 \quad (9.15)$$

$$\int_0^\lambda \cos(akx) \cos(bkx) dx = \frac{\lambda}{2} \delta_{ab} \quad (9.16)$$

$$\int_0^\lambda \sin(akx) \sin(bkx) dx = \frac{\lambda}{2} \delta_{ab} \quad (9.17)$$

where δ_{ab} is the Kronecker delta.

To find the coefficients, then, we multiply both sides of equation (9.11) by $\cos(lkx)$ where l is a positive integer. Then we integrate over one wavelength.

$$\int_0^\lambda f(x) \cos(lkx) dx = \int_0^\lambda \frac{A_o}{2} \cos(lkx) dx \quad (9.18)$$

$$+ \int_0^\lambda \sum_{m=1}^{\infty} A_m \cos(mkx) \cos(lkx) dx \quad (9.19)$$

$$+ \int_0^\lambda \sum_{m=1}^{\infty} B_m \sin(mkx) \cos(lkx) dx \quad (9.20)$$

which gives

$$\int_0^\lambda f(x) \cos(mkx) dx = \int_0^\lambda A_m \cos(mkx) \cos(mkx) dx \quad (9.21)$$

that is, only the term with two cosine functions where $l = m$ will be non zero. So

$$\int_0^\lambda f(x) \cos(mkx) dx = \frac{\lambda}{2} A_m \quad (9.22)$$

solving for A_m we have

$$A_m = \frac{2}{\lambda} \int_0^\lambda f(x) \cos(mkx) dx \quad (9.23)$$

We can perform the same steps to find B_m only we use $\sin(lkx)$ as the multiplier. Then we find

$$B_m = \frac{2}{\lambda} \int_0^\lambda f(x) \sin(mkx) dx \quad (9.24)$$

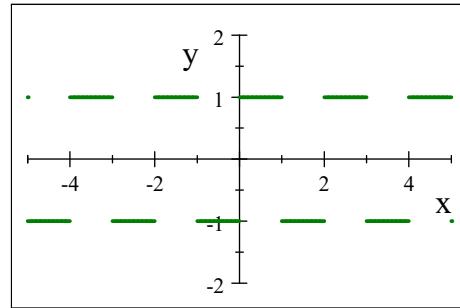
Square wave

Let's find the series for a square wave using our Fourier analysis technique.

Let's take

$$\lambda = 2 \quad (9.25)$$

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < \frac{\lambda}{2} \\ -1 & \text{if } \frac{\lambda}{2} < x < \lambda \end{cases} \quad (9.26)$$



since $f(x)$ is odd, $A_m = 0$ for all m . We have

$$B_m = \frac{2}{\lambda} \int_0^{\frac{\lambda}{2}} (1) \sin(mkx) dx + \frac{2}{\lambda} \int_{\frac{\lambda}{2}}^\lambda (-1) \sin(mkx) dx \quad (9.27)$$

so

$$B_m = \frac{1}{m\pi} (-\cos(mkx)|_0^{\frac{\lambda}{2}} + \frac{1}{m\pi} (\cos(mkx)|_{\frac{\lambda}{2}}^\lambda) \quad (9.28)$$

Which is

$$B_m = \frac{1}{m\pi} \left(1 \cos\left(m\frac{2\pi}{\lambda}x\right)|_0^{\frac{\lambda}{2}} + \frac{1}{m\pi} \left(\cos\left(m\frac{2\pi}{\lambda}x\right)\right|_{\frac{\lambda}{2}}^\lambda \right) \quad (9.29)$$

so

$$B_m = \frac{1}{m\pi} \left(\left(-\cos \left(m \frac{2\pi}{\lambda} \frac{\lambda}{2} \right) \right) + \cos \left(m \frac{2\pi}{\lambda} (0) \right) \right) \quad (9.30)$$

$$+ \frac{1}{m\pi} \left(\left(\cos \left(m \frac{2\pi}{\lambda} \lambda \right) - \cos \left(m \frac{2\pi}{\lambda} \frac{\lambda}{2} \right) \right) \right) \quad (9.31)$$

which is

$$B_m = \frac{2}{m\pi} (1 - \cos(m\pi)) \quad (9.32)$$

Our series is then just

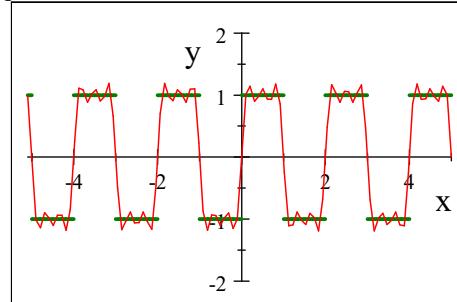
$$f(x) = \sum_{m=1}^{\infty} \frac{2}{m\pi} (1 - \cos(m\pi)) \sin(mkx) \quad (9.33)$$

and we can write a few terms

Term	
1	$\frac{4}{\pi} \sin(kx)$
2	0
3	$\frac{4}{3\pi} \sin(3kx)$
4	0
5	$\frac{4}{5\pi} \sin(5kx)$

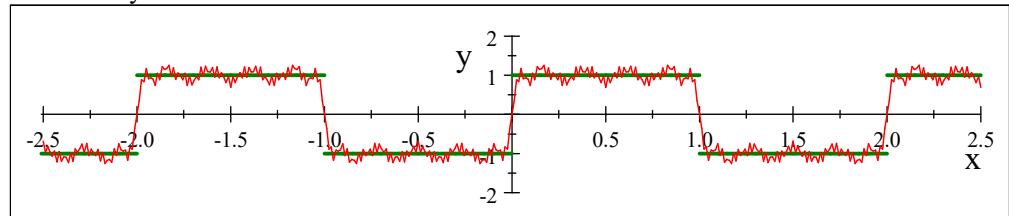
(9.34)

then the partial sum up to $m = 5$ looks like



$$f(x) = \frac{4}{\pi} \sin(kx) + \frac{4}{3\pi} \sin(3kx) + \frac{4}{5\pi} \sin(5kx) \quad (9.35)$$

With twenty terms we would have



In the limit of infinitely many waves, the match would be perfect. But we don't usually

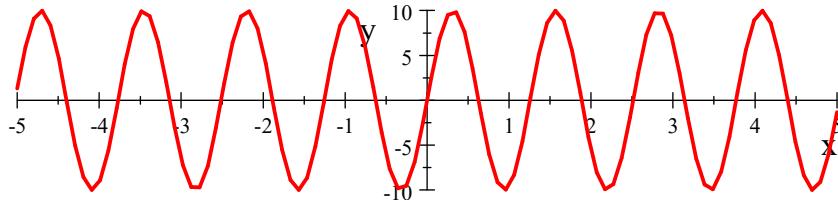
need an infinite number of terms. we can pick the part of the spectrum that best represents the phenomena we desire to observe. For example, oil based compounds all have specific spectral signatures in the wavelength range between 3 – 5 micrometers. If you wish to tell the difference between gasoline and crude oil, you can restrict your study to these wavelengths alone.

Frequency Uncertainty for Signals and Particles

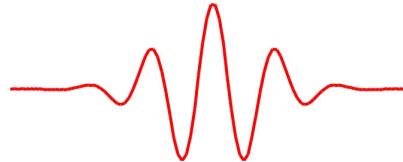
Up to this lecture, when we thought of a wave we have mostly thought of something like this

$$y = y_{\max} \sin(kx - \omega t - \phi)$$

which in practice might look like this



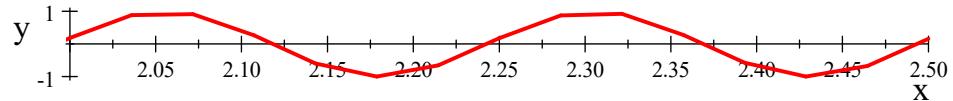
And we have noticed that there is no start or stop to this kind of wave. Our figure starts at $x = -5$ and ends at $x = 5$, but the equation does not! There is a value of y for every x from $-\infty$ to $+\infty$. But many signals are not such waves. They may be very limited in size.



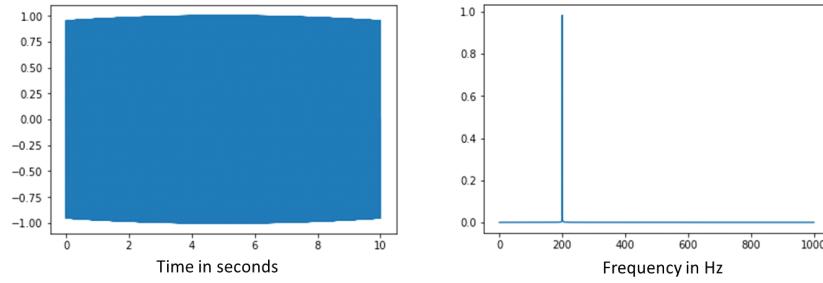
We should investigate what happens when you have a limited wave. I did this investigation using Python. Suppose we have a sine wave with $f = 200$ Hz, but I limit this wave's existence by making it start at $t_i = 0$ and then make it end at $t_f = 10$ s. I could do this in practice by turning on a radio transmission or even an acoustic speaker, and then turning the device off ten seconds later. Our screen resolution is terrible for plotting such a function, but in the figure below you can see that our signal only exists from $t = 0$ to $t = 10$ s.



If I zoom in on a part of the graph, we can see that it is really a sin wave.



Python did equally bad at plotting this. All we see is a blue band.



10s, 200Hz

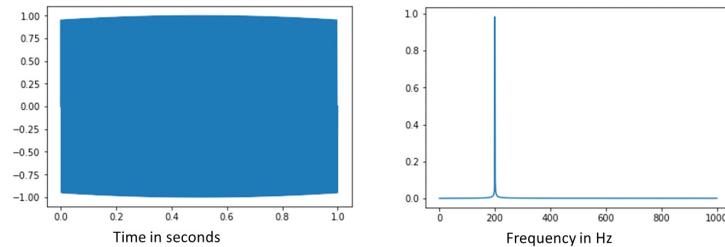
But in the second graph, notice that we have plotted frequency. Python and most scientific programming languages have functions to find a digital version of a Fourier transform to produce a spectrum. It takes in a signal and finds which frequencies are in a signal. It performs the job of a spectrometer, so we would call the figure to the right a spectrogram (or just a spectrum).

We expect only one frequency, 200 Hz, and that is mostly what we get. Since our period for our wave is

$$T = \frac{1}{200 \text{ Hz}} = 0.005 \text{ s}$$

and we have 10 s of data, that is four orders of magnitude more signal than a period.

The whole signal seems very long compared to a period. We expect this to look kind of like an infinite signal. But suppose we take the same wave, but for less time. We limit the wave more.

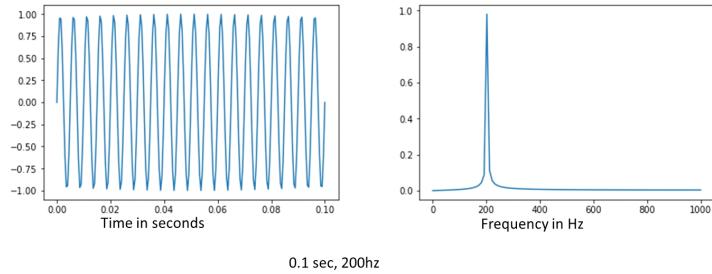


1sec, 200hz

So we still get a blue blur for our wave picture, but now the wave only exists for one

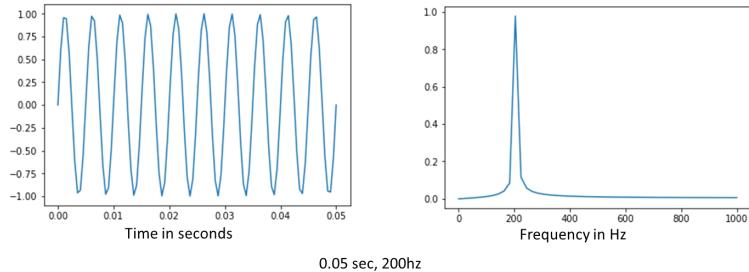
instead of ten seconds. If you look closely at the frequency graph, you will notice that the 200 Hz peak representing our wave is a bit wider right at the bottom.

We could limit our wave more, say, so it only lasts $t_f = 0.1$ s. We would get a set of graphs that look like this.

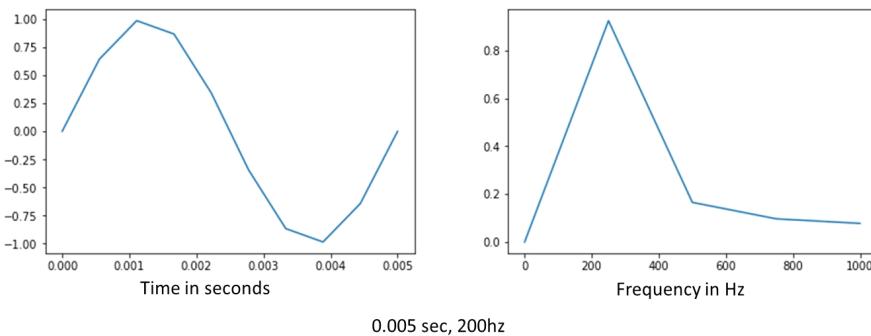


0.1 sec, 200hz

Notice that not only can Python render the wave now, but more importantly the 200 Hz frequency peak is noticeable wider. This is profound! It means that by limiting the wave, we no longer have just one frequency! The graph tells us we have mostly 200 Hz but we also have some 199 Hz and some 201 Hz and some 190 Hz and some 210 Hz, etc. The very fact that the wave does not go on so long requires that we have more than one frequency in the wave. We could say that as Δt gets smaller, our Δf is getting bigger. Here are two more examples with smaller Δt values.



0.05 sec, 200hz



The cost of limiting our waves is that we can't have a single frequency for the wave. For an engineer, this means that if you only measure a short segment of the signal, you have an increased uncertainty in the frequency you will find from that signal. For chemists, it means that when we look at quantum wave functions we can expect uncertainty in the frequency (or wavelength) because the quantum particle (like an electron) is limited. It is important to know what we mean by uncertainty in this case. In our example above, when we say that Δf increased we really mean that we have more than one frequency. We don't just mean that we don't know the frequency well. We really are mixing more than one frequency. This idea of limiting waves creating uncertainty shows up in physical chemistry as Heisenberg's uncertainty principle.

10 Interference of light waves

Fundamental Concepts

- Light is a wave in the electromagnetic field
- Light is a superposition of many small waves called photons
- The energy in a photon is proportional to the frequency of the photon
- If we mix two coherent light sources, we get interference, with an intensity pattern given by $I = I_{\max} \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right)$
- Most detectors cannot follow the fluctuation of light because their integration time is too long.

The Nature of Light

Physical Ideas of the nature of Light

Before the 19th century (1800's) light was assumed to be a stream of particles. Newton was the chief proponent of this theory. The theory was able to explain reflection of light from mirrors and other objects and therefore explain vision. In 1678 Huygens showed that wave theory could also explain reflection and vision.

In 1801 Thomas Young demonstrated that light had attributes that were best explained by wave theory. We will study Young's experiment later today. The crux of his experiment was to show that light displayed constructive and destructive interference—clearly a wave phenomena! The theory of the nature of light took a dramatic shift

In 1805 Joseph Smith was born in Sharon, Vermont.

In September of 1832 Joseph Smith received a revelation that said in part :

For the word of the Lord is truth, and whatsoever is truth is light, and

whatsoever is light is Spirit, even the Spirit of Jesus Christ. And the Spirit giveth light to every man that cometh into the world; and the Spirit enlighteneth every man through the world, that hearkeneth to the voice of the Spirit. (D&C 84:45-46)

In December of 1832 Joseph Smith received another revelation that says in part:

This Comforter is the promise which I give unto you of eternal life, even the glory of the celestial kingdom; which glory is that of the church of the Firstborn, even of God, the holiest of all, through Jesus Christ his Son—He that ascended up on high, as also he descended below all things, in that he comprehended all things, that he might be in all and through all things, the light of truth; which truth shineth. This is the light of Christ. As also he is in the sun, and the light of the sun, and the power thereof by which it was made. As also he is in the moon, and is the light of the moon, and the power thereof by which it was made; as also the light of the stars, and the power thereof by which they were made; and the earth also, and the power thereof, even the earth upon which you stand. And the light which shineth, which giveth you light, is through him who enlighteneth your eyes, which is the same light that quickeneth your understandings; which light proceedeth forth from the presence of God to fill the immensity of space—the light which is in all things, which giveth life to all things, which is the law by which all things are governed, even the power of God who sitteth upon his throne, who is in the bosom of eternity, who is in the midst of all things. (D&C 88:5-12)

Light, even real, physical light, seems to be of interest to Latter Day Saints.

In 1847 the saints entered the Salt Lake Valley.

In 1873 Maxwell published his findings that light is an electromagnetic wave (something we will try to show before this course is over!).

Planck's work in quantization theory (1900) was used by Einstein In 1905 to give an explanation of the photoelectric effect that again made light look like a particle.

Current theory allows light to exhibit the characteristics of a wave in some situations and like a particle in others. We will study both before the end of the semester.

The results of Einstein's work give us the concept of a *photon* or a quantized unit of radiant energy. Each “piece of light” or photon has energy

$$E = hf \quad (10.1)$$

where f is the frequency of the light and h is a constant

Question 223.10.1

$$h = 6.63 \times 10^{-34} \text{ Js} \quad (10.2)$$

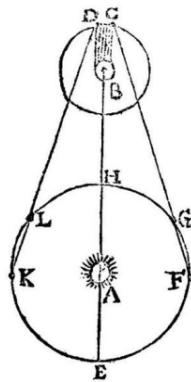
Question 223.10.2

The nature of light is fascinating and useful both in physical and religious areas of thought.

Measurements of the Speed of Light

One of the great foundations of modern physical theory is that the speed of light is constant in a vacuum. Galileo first tried to measure the speed of light. He used two towers in town and placed a lantern and an assistant on each tower. The lanterns had shades. The plan was for one assistant to remove his shade, and then for the assistant on the other tower to remove his shade as soon as he saw the light from the first lantern. Back at the first tower, the first assistant would use a clock to determine the time difference between when the first lantern was un-shaded, and when they saw the light from the second tower. The light would have traveled twice the inter-tower distance. Dividing that distance by the time would give the speed of light. You can probably guess that this did not work. Light travels very quickly. The clocks of Galileo's day could not measure such a small time difference. Ole Rømer was the first to succeed in measuring the speed of light.

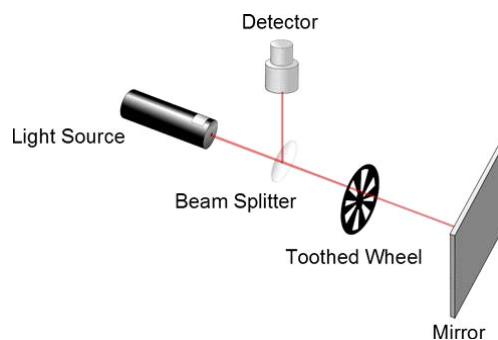
Rømer's Measurement of the speed of light



A diagram illustrating Rømer's determination of the speed of light. Point A is the Sun, point B is Jupiter. Point C is the immersion of Io into Jupiter's shadow at the start of an eclipse

Rømer performed his measurement in 1675, 269 years before digital devices existed!. He used the period of revolution of Io, a moon of Jupiter, as Jupiter revolved around the sun. He first measured the period of Io's rotation about Jupiter, then he predicted an eclipse of Io three months later. But he found his calculation was off by 600 s. After careful thought, he realized that the Earth had moved in its orbit, and that the light had to travel the extra distance due to the Earth's new position. Given Rømer's best estimate for the orbital radius of the earth and his time difference, Rømer arrived at a estimate of $c = 2.3 \times 10^8 \frac{\text{m}}{\text{s}}$. Amazing for 1675!

Fizeau's Measurement of the speed of light



Hippolyte Fizeau measured the speed of light in 1849 using a toothed wheel and a mirror and a beam of light. The light passed through the open space in the wheel's teeth

as the wheel rotated. Then was reflected by the mirror. The speed would be

$$v = \frac{\Delta x}{\Delta t}$$

We just need Δx and Δt .

It is easy to see that

$$\Delta x = 2d$$

because the light travels twice the distance to the mirror (d) and back. So the speed is just

$$v = \frac{2d}{\Delta t}$$

If the wheel turned just at the right angular speed, then the reflected light would hit the next tooth and be blocked. Think of angular speed

$$\omega = \frac{\Delta\theta}{\Delta t}$$

so the time difference would be

$$\Delta t = \frac{\Delta\theta}{\omega}$$

we find $\Delta\theta$ by taking the number of teeth on the wheel and dividing by 2π by that number.

Then the speed of light must be

$$\begin{aligned} c &= v = \frac{2d}{\frac{\Delta\theta}{\omega}} \\ &= \frac{2d\omega}{\Delta\theta} \\ &= \frac{2d\omega N_{teeth}}{2\pi} \\ &= \frac{d\omega N_{teeth}}{\pi} \end{aligned}$$

then if we have 720 teeth and ω is measured to be $d = 7500$ m

$$\begin{aligned} c &= \frac{(7500 \text{ m})(172.79 \text{ Hz})(720)}{\pi} \\ &= 2.97 \times 10^8 \frac{\text{m}}{\text{s}} \end{aligned}$$

which is Fizeau's number and it is pretty good!

Modern measurements are performed in very much the same way that Fizeau did his calculation. The current value is

$$c = 2.9979 \times 10^8 \frac{\text{m}}{\text{s}} \quad (10.3)$$

Faster than light

Pass the photon
(ball) demo

Question 223.10.3

Question 223.10.4

The speed of light in a vacuum is constant, but in matter the speed of light changes. We will study this in detail when we look at refraction. But for now, a dramatic example is Cherenkov radiation. It is an eerie blue glow around the core of nuclear reactors. It occurs when electrons are accelerated past the speed of light in the water surrounding the core. The electrons emit light and the light waves form a Doppler cone or a light-sonic boom! The result is the blue glow.

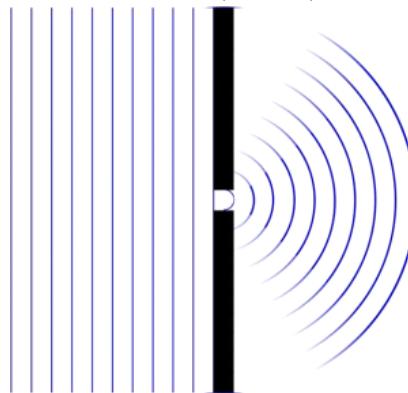


Cherenkov radiation from a 250kW TRIGA reactor. (Image in the Public Domain, courtesy US Department of Energy)

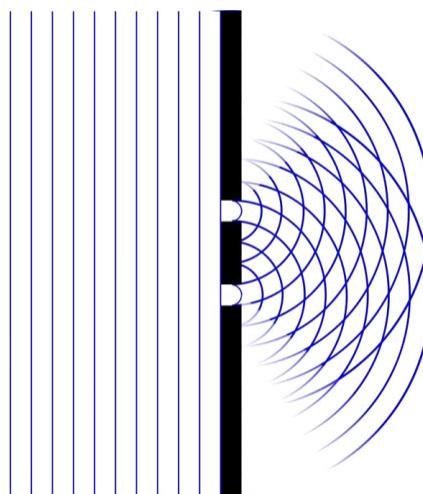
This does bring up a problem in terminology. What does the word “medium” mean? We have used it to mean the substance through which a wave travels. This substance must have the property of transferring energy between its parts, like the coils of a spring can transfer energy to each other, or like air molecules can transfer energy by collision. For light the wave medium is the electromagnetic field. This field can store and transfer energy (we will see this later in the course). But many books on physics call materials like glass a “medium” through which light travels. The water in our last example is such a medium. Are glass and water wave mediums for light? The answer is no. Light does not need any matter to form its wave. The wave medium is the electromagnetic field. So we will have to keep this in mind as we allow light to travel through matter. We may call the matter a “medium,” but it is not the wave medium.

Interference and Young's Experiment

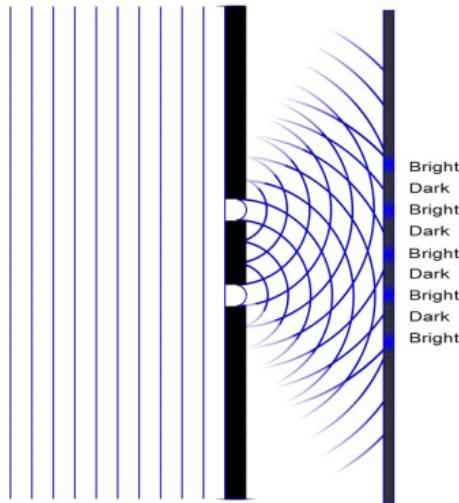
Waves do some funny things when they encounter barriers. Think of a water wave. If we pass the wave through a small opening in a barrier, the wave can't all get through the small hole, but it can cause a disturbance. We know that a small disturbance will cause a wave. But this wave will be due to a very small—almost point—source. So the waves will be spherical leaving the opening. The smaller the opening the more pronounced the curving of the wave, because the source (the hole) is more like a point source.



Now suppose we have two of these openings. We expect the two sources to make curved waves and those waves can interfere.

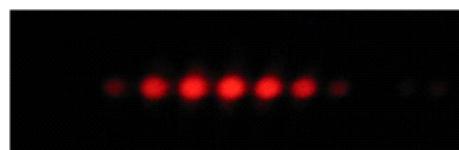


In the figure, we can already see that there will be constructive and destructive interference were the waves from the two holes meet. Thomas young predicted that we should see constructive and destructive interference in light (he drew figures very like the ones we have used to explain his idea).



Young's Experiment demo

Young set up a coherent source of light and placed it in front of this source a barrier with two very thin slits cut in it to test his idea.. He set up a screen beyond the barrier and observed the pattern on the screen formed by the light. This (in part) is what he saw



Question 223.10.5

We see bright spots (constructive interference) and dark spots (destructive interference). Only wave phenomena can interfere, so this is fairly good evidence that light is a wave.

Constructive Interference

We can find the condition for getting a bright or a dark band if we think about it a bit. Here are our equations that we developed for constructive and destructive interference.

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = m2\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad \text{Constructive}$$

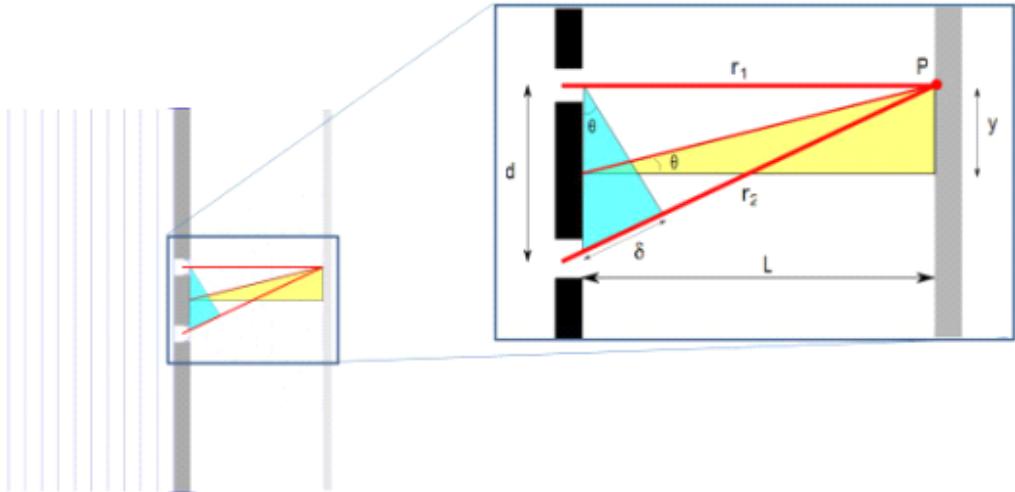
$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = (2m + 1)\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad \text{Destructive}$$

For constructive interference, the difference in phase, $\Delta\phi$, must be a multiple of 2π .

That means the path difference between the two slit-sources must be an even number of wavelengths. We have been calling the path difference in the total phase Δx , or for

spherical waves Δr , but in optics it is customary to call this path difference δ . So

$$\delta = \Delta r$$



and our total phase equation becomes

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \delta + \Delta\phi_o \right) = m2\pi$$

Our light going through the slits is all coming from the same light source. So as long as the light hits the slits at a 90° angle, $\Delta\phi_o = 0 - 0 = 0$ so we don't have a change in phase constant. But we do have a change in $\Delta r = \delta$. Let's suppose that the screen is far away so the distance from the slits to the screen, $L \gg d$, the slit distance. Then we can say that the blue triangle is almost a right triangle, and then δ is

$$\delta = r_2 - r_1 \approx d \sin \theta$$

so then

$$\Delta\phi = \left(\frac{2\pi}{\lambda} d \sin \theta + 0 \right) = m2\pi$$

We can do a little math to make this simpler.

$$\begin{aligned} \frac{2\pi}{\lambda} d \sin \theta &= m2\pi \\ \frac{1}{\lambda} d \sin \theta &= m \\ d \sin \theta &= m\lambda \end{aligned}$$

We started by knowing our wave needs to sift by an integer number times 2π radians but now we see that is equivalent to shifting an integer number times the wavelength, λ . This will make the two waves experience constructive interference (a bright spot).

$$\delta = d \sin \theta = m\lambda \quad (m = 0, \pm 1, \pm 2, \dots) \quad \text{Constructive}$$

Question 223.10.6

where in optics m is called the *order number*. That is, if the two waves are off by any

number of whole wavelengths then our total phase due to path difference will be 2π . In optics, the bright spots formed by constructive interference are called *fringes*.

If we assume that $\lambda \ll d$ we can find the distance from the axis for each fringe more easily. This condition guarantees that θ will be small. Using the yellow triangle we see

$$\tan \theta = \frac{y}{L}$$

but if θ is small this is just about the same as

$$\sin \theta = \frac{y}{L}$$

because for small angles $\tan \theta \approx \sin \theta \approx \theta$. So if theta is small then

$$\begin{aligned}\delta &= d \sin \theta \\ &= d \frac{y}{L}\end{aligned}$$

and for a bright spot or fringe we find

$$d \frac{y}{L} = m\lambda$$

Solving for the position of the bright spots gives

$$y_{bright} = \frac{\lambda L}{d} m \quad (m = 0, \pm 1, \pm 2, \dots) \quad (10.4)$$

We can measure up from the central spot and predict where each successive bright spot will be.

Destructive Interference

We can also find a condition for destructive interference. Our destructive interference equation is

$$\Delta\phi = \left(\frac{2\pi}{\lambda} \Delta r + \Delta\phi_o \right) = (2m + 1)\pi$$

Once again $\Delta\phi_o = 0$ and $\Delta r = \delta$

$$\begin{aligned}\left(\frac{2\pi}{\lambda} \delta \right) &= (2m + 1)\pi \\ \left(\frac{2}{\lambda} \delta \right) &= (2m + 1) \\ \delta &= \frac{\lambda}{2} (2m + 1) \\ \delta &= \lambda \left(m + \frac{1}{2} \right)\end{aligned}$$

This just shows us again that a path difference of an odd multiple of a half wavelength will give distractive interference.

$$\delta = d \sin \theta = \left(m + \frac{1}{2} \right) \lambda \quad (m = 0, \pm 1, \pm 2, \dots)$$

will give a dark fringe. The location of the dark fringes will be

$$y_{dark} = \frac{\lambda L}{d} \left(m + \frac{1}{2} \right) \quad (m = 0, \pm 1, \pm 2 \dots) \quad (10.5)$$

Double Slit Intensity Pattern

The fringes we have seen are not just points, but are patterns that fade from a maximum intensity. This is why they are called fringes. We can calculate the intensity pattern to show this. We need to know a little bit about electric fields to do this.

Electric field preview

We can represent an electromagnetic wave using just the electric field (the magnetic field pattern is very similar and can be derived from the electric field pattern).

We represent the field by an equation like

$$y = y_0 \sin (kr - \omega t)$$

but since the medium for light waves is the electric field, let's use the symbol E instead of y so we can see that we have a change in the field strength and not a displacement of some material thing.

$$E = E_{\max} \sin (kr - \omega t) \quad (10.6)$$

where the amplitude of the wave is E_{\max} and ω is the angular frequency. This is just our traveling wave equation, but with electric field strength, labeled E , for the amplitude.

Then to find the intensity pattern, we take two waves in the electric field, one from slit one

$$E_1 = E_{\max} \sin (kr_1 - \omega t + \phi_o)$$

and the other from slit two.

$$E_2 = E_{\max} \sin (kr_2 - \omega t + \phi_o)$$

This is mathematically just like superposition of sound waves.

Superposition of two light waves

Remember when we superimposed waves before, we mixed the waves

$$y_1 = A \sin (kr_1 - \omega t + \phi_1)$$

$$y_2 = A \sin (kr_2 - \omega t + \phi_2)$$

and using

$$\sin a + \sin b = 2 \cos\left(\frac{a-b}{2}\right) \sin\left(\frac{a+b}{2}\right)$$

we found the resultant wave

$$y_r = 2A \cos\left(\frac{1}{2}(\Delta\phi)\right) \sin\left(k\frac{r_2+r_1}{2} - \omega t + \frac{\phi_2+\phi_1}{2}\right)$$

Our light waves are just two waves. They may be the superposition of many individual photons, but the combined wave is just a wave.

At the slits, the waves have the same amplitude E_{\max} and the same phase constant, $\phi_1 = \phi_2 = \phi_o$, but E_2 travels farther than E_1 , so $\Delta\phi$ is due to the path difference. We expect to find that the path difference would be

$$\begin{aligned}\Delta\phi &= k\Delta r + \Delta\phi_o \\ &= k\delta + 0 \\ &= \frac{2\pi}{\lambda}d \sin\theta\end{aligned}$$

Now superimposing E_1 and E_2 at point P on the screen gives

$$\begin{aligned}E_P &= E_2 + E_1 \\ &= E_{\max} \sin(kr_2 - \omega t) + E_o \sin(kr_1 - \omega t)\end{aligned}$$

and using our prior result, we have

$$E_P = 2E_{\max} \cos\left(\frac{1}{2}\Delta\phi\right) \sin\left(k\frac{(r_2+r_1)}{2} - \omega t + \phi_o\right)$$

and using our equation for $\Delta\phi$ above we get

$$E_P = 2E_{\max} \cos\left(\frac{1}{2}\left(\frac{2\pi}{\lambda}d \sin\theta\right)\right) \sin\left(k\frac{(r_2+r_1)}{2} - \omega t + \phi_o\right)$$

We have a combined wave at point P that is a traveling wave $\left(\sin\left(k\frac{(r_2+r_1)}{2} - \omega t + \phi_o\right)\right)$ but with amplitude $\left(2E_{\max} \cos\left(\frac{1}{2}\left(\frac{2\pi}{\lambda}d \sin\theta\right)\right)\right)$ that depends on our total phase $\Delta\phi = \frac{2\pi}{\lambda}d \sin\theta$.

But the situation is more complicated because of how we detect light. Our eyes, and most detectors measure the intensity of the light. We know that

$$I = \frac{\mathcal{P}}{A}$$

later in the course we will show that the power in an electromagnetic field wave is proportional to the square of the electric field displacement.

$$\mathcal{P} \propto E_P^2 \tag{10.7}$$

For now, let's just assume this is true. Then the intensity must be proportional to the amplitude of the electric field squared.

$$\begin{aligned} I &\propto E_P^2 \\ &= 4E_{\max}^2 \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right) \sin^2 \left(\frac{k(r_2 + r_1)}{2} - \omega t + \phi_o \right) \end{aligned}$$

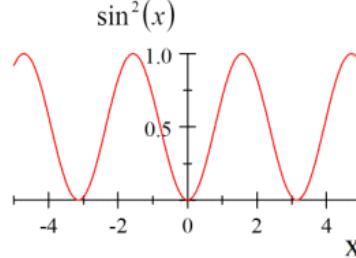
Light detectors collect energy for a set amount of time. So most light detection will be a value averaged over a set *integration time*. This means that the detector sums up (or integrates) the amount of power received over the detector time. Usually the integration time is much longer than a period, so what is really detected is like a time-average of our intensity.

$$\begin{aligned} \int_{\text{many T}} Idt &\propto = \int_{\text{many T}} 4E_{\max}^2 \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right) \sin^2 \left(\frac{k(r_2 + r_1)}{2} - \omega t + \phi_o \right) dt \\ &= 4E_{\max}^2 \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right) \int_{\text{many T}} \sin^2 \left(\frac{k(rx_2 + r_1)}{2} - \omega t + \phi_o \right) dt \end{aligned}$$

but the term

$$\int_{\text{many T}} \sin^2 \left(\frac{k(r_2 + r_1)}{2} - \omega t + \phi_o \right) dt = \frac{1}{2} \quad (10.8)$$

To convince yourself of this, think that $\sin^2(\omega t)$ has a maximum value of 1 and a minimum of 0. Looking at the graph



should be convincing that the average value over a period is 1/2. The average over many periods will still be 1/2.

So we have

$$\bar{I} = \int_{\text{many periods}} Idt \propto 2E_{\max}^2 \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right) \quad (10.9)$$

where \bar{I} is the time average intensity. The important part is that the time varying part has averaged out.

So, usually in optics, we ignore the fast fluctuating parts of such calculations because we can't see them and so we write

$$I = I_{\max} \cos^2 \left(\frac{1}{2} \left(\frac{2\pi}{\lambda} d \sin \theta \right) \right)$$

where we have dropped the bar from the I , but it is understood that the intensity we report is a time average over many periods.

We should remind ourselves, of our intensity pattern

$$I = I_{\max} \cos^2 \left(\frac{1}{2} \frac{2\pi}{\lambda} d \sin \theta \right)$$

is really

$$I = I_{\max} \cos^2 \left(\frac{\Delta\phi}{2} \right)$$

Which is just our amplitude squared for the mixing of two waves. All we have done to find the intensity pattern is to find and expression for the phase difference $\Delta\phi$.

Our intensity pattern should give the same location for the center of the bright spots as we got before. Let's check that it works. We used the small angle approximation before, so let's use it again now. For small angles

$$\begin{aligned} I &= I_{\max} \cos^2 \left(\frac{\pi d}{\lambda} \theta \right) \\ &= I_{\max} \cos^2 \left(\frac{\pi d}{\lambda} \frac{y}{L} \right) \end{aligned}$$

Then we have constructive interference when

$$\frac{\pi d}{\lambda} \frac{y}{L} = m\pi$$

or

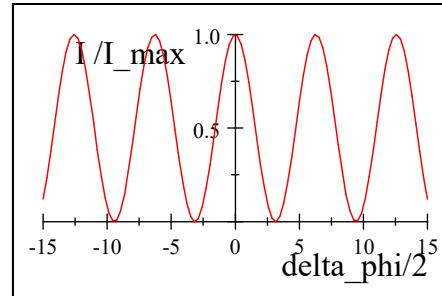
$$y = m \frac{L\lambda}{d}$$

which is what we found before.

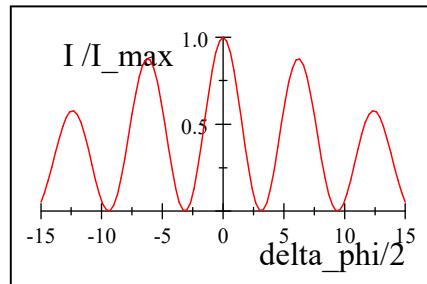
The plot of normalized intensity

$$\frac{I}{I_{\max}} = \cos^2 \left(\frac{\Delta\phi}{2} \right)$$

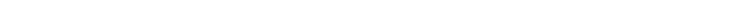
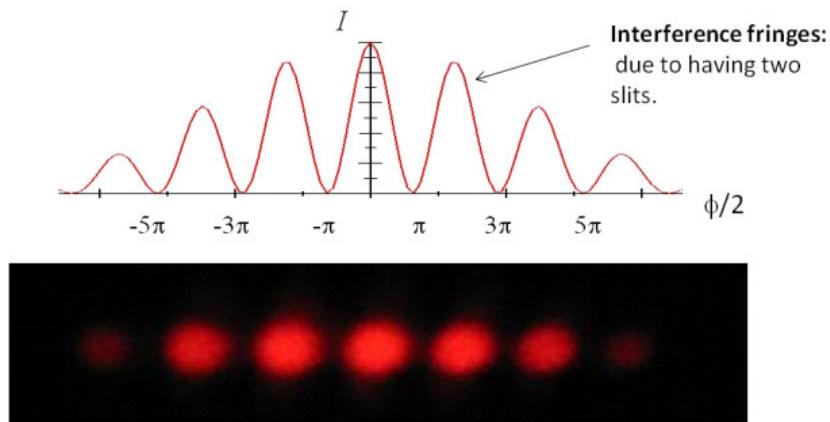
verses $\Delta\phi/2$ is given next,



but we will find that we are not quite through with this analysis. Next time we will find that there is another compounding factor that reduces the intensity as we move away from the midpoint.



Let's pause to remember what this pattern means. This is the intensity of light due to interference. It is instructive to match our intensity pattern that Young saw with our graph.



The high intensity peaks are the bright fringes and the low intensity troughs are the dark fringes. The pattern moves smoothly and continuously from bright to dark.

11 Many slits, and single slits

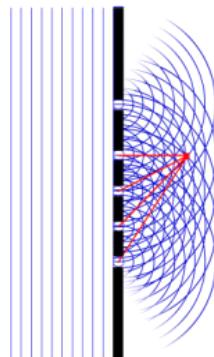
Last lecture we found the pattern that results from sending light through two slits. This lecture takes on many slits, and even the pattern that results from a single slit.

Fundamental Concepts

- Many slit devices are called diffraction gratings
- These devices can be used to build spectrometers
- Single slits also produce an interference pattern

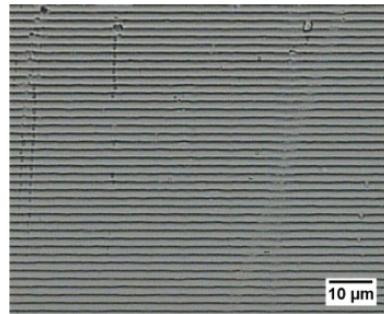
Diffraction Gratings

We have discussed the interference that comes from having two small slits. But what if we have more slits?



Rainbow Glasses

A diffraction grating is an optical element with many parallel slits spaced very close together. Here is a typical diffraction grating created by etching lines in a piece of glass. The etchings scatter the light, but the un-etched part allows the light to pass through. The un-etched parts are essentially a series of slits.



Surface of a diffraction grating (600 lines/mm). Image taken with optical transmission microscope. (Image in the public domain courtesy Scapha)

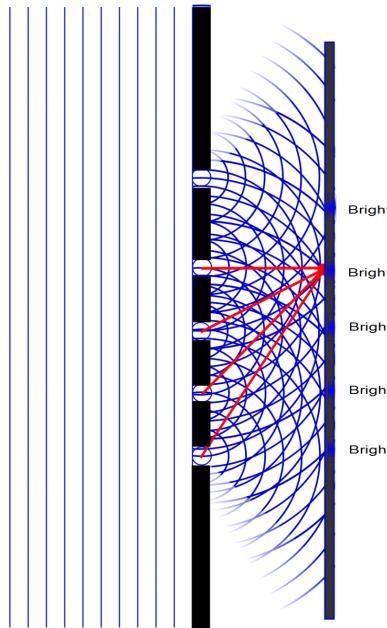
A typical grating might have 5000 slits per unit centimeter. You have probably used a diffraction grating to see rainbow colors in a beginning science class.

If we use $5000 \frac{\text{slits}}{\text{cm}}$ for an example, we see that the slit spacing is

$$d = \frac{1}{5000} \text{ cm} \quad (11.1)$$

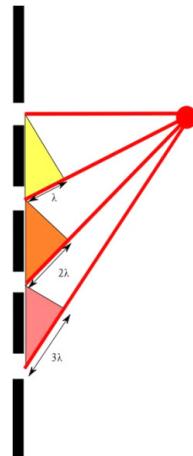
$$= 2.0 \times 10^{-6} \text{ m} \quad (11.2)$$

Take a section of diffraction grating as shown below



At some point, two of the slits will have a path difference that is a whole wavelength, and we would expect a bright spot. But what about the other slits? If we have a

slit spacing such that each of the succeeding slits has a path difference that is just an additional wavelength, then each of the slits will contribute to the constructive interference at our point, and the point will become a very bright spot.

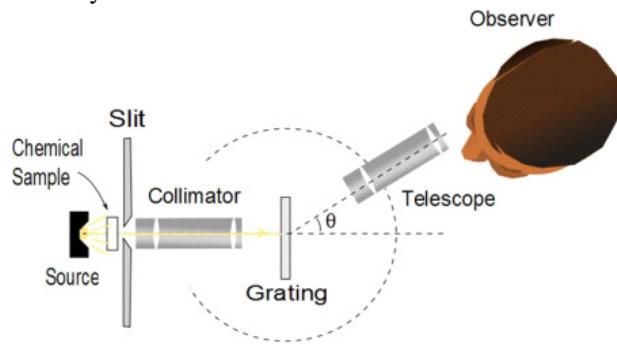


The light leaves each slit in phase with the light from the rest of the slits. At some distance L away and at some angle θ , we will have a path difference

$$\delta = d \sin(\theta_{bright}) = m\lambda \quad m = 0, \pm 1, \pm 2, \dots \quad (11.3)$$

This looks a lot like our condition for constructive interference for two slits.

This equation tells us that each wavelength, λ , will experience constructive interference at a slightly different angle θ_{bright} . Knowing d and θ allows an accurate calculation of λ . This may seem a silly thing to do, but suppose we add into our system a sample of a chemical to identify



We could then record the intensity of the transmitted light as a function of angle, which is equivalent to λ . We can again generate a spectrum. This is a traditional way to build

Demo a student spectrometer with a gas tube a spectrometer and many such devices are available today.

Resolving power of diffraction gratings

For two nearly equal wavelengths λ_1 and λ_2 , we say that the diffraction grating can resolve the wavelengths if we can distinguish the two using the grating. The *resolving power* of the grating is defined as

$$R = \frac{(\lambda_1 + \lambda_2)}{2(\lambda_1 - \lambda_2)} = \frac{\bar{\lambda}}{\Delta\lambda} \quad (11.4)$$

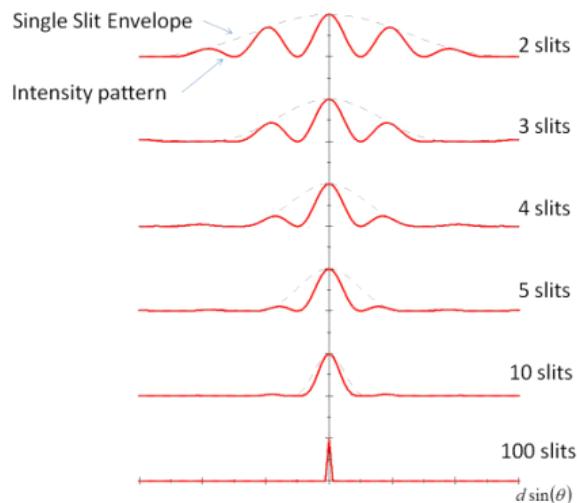
We can show that for the m -th order diffraction, the resolving power is

$$R = Nm \quad (11.5)$$

where N is the number of slits. So our ability to distinguish wavelengths increases with the number of slits and with the order (which is related to how far off-axis we look).

Note that for $m = 0$ we have no ability to resolve wavelengths. The central peak is a mix of all wavelengths and usually looks white for normal illumination.

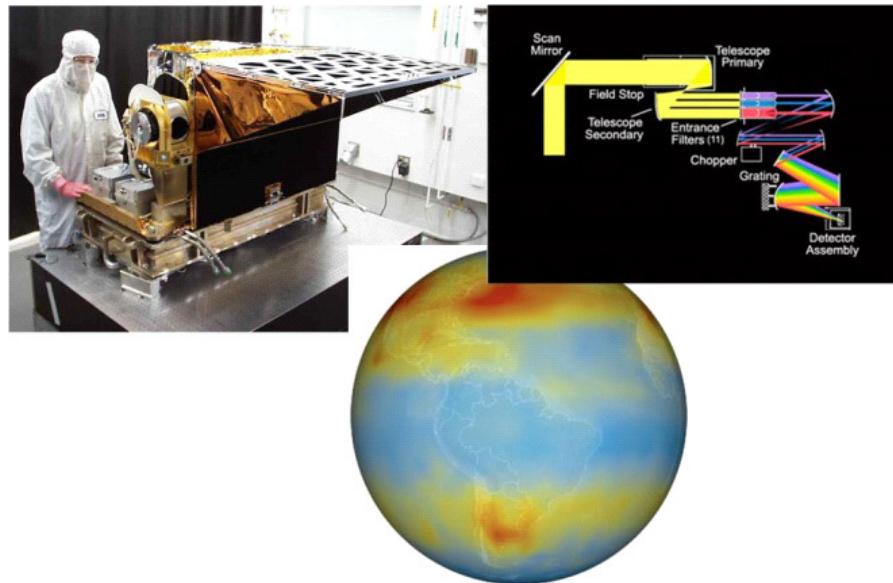
That the resolution depends on the number of slits, N , means that we can improve our spectrometer by using more lines. Here is a representation of what happens as we increase N



we can see that the peaks get narrower as N increases. These graphs are for a particular λ . If the peaks for a particular λ get narrower, then there will be less overlap with adjacent λ 's which means that each wavelength can more easily be resolved.

Spectrometers are used in many places. One that has some public interest today is

monitoring the atmosphere. Instruments like the one shown below detect the amount of special gasses in the atmosphere using IR spectrometers.



AIRS sensor, spectrometer design, and global CO₂ data. (Images in the Public Domain courtesy NASA)

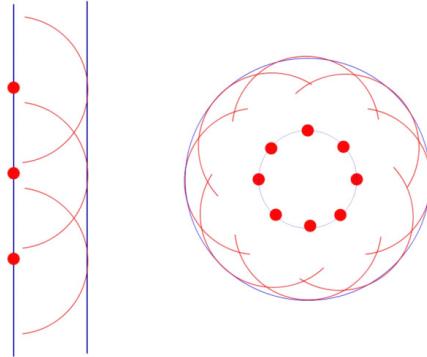
The instrument shown is the AIRS spectrometer. You can see in the diagram that it uses a grating spectrometer. The picture of the Earth is a composite of AIRS data showing the northern and southern bands of CO₂.

Single Slits

We have looked at interference from two slits, and for many slits. The two slits acted like two coherent sources. We might expect that a single slit will give only a single bright spot. But let's consider a single slit very closely. To do this, let's return to the work of Huygens.⁸ His idea for the nature of light was simple. He suggested that every point on the wave front of a light wave was the source (the disturbance) for a new set of small spherical waves. The next wavefront would be formed by the superposition of the

⁸ Huygens method is technically not a correct representation of what happens. The actual wave leaving the single opening is a superposition of the original wave, and the wave scattered from the sides of the opening. You can see this scattering by tearing a small hole in a piece of paper and looking through the hole at a light source. You will see the bright ring around the hole where the edges of the paper are scattering the light. But the mathematical result we will get using Huygens method gives a mathematically identical result for the resulting wave leaving the slit with much less high power math. So we will stick with Huygens in this class.

little “wavelets.” Here is an example for a plane wave and a spherical wave.



In each case we have drawn spots on the wave front and drawn spherical waves around those spots. where the wavefronts of the little wavelets combine, we have new wave front of our wave. This is sort of what happens in bulk matter. Remember that light is absorbed and re-emitted by the atoms of the material. This is why light slows down in a medium. Because of the time it is absorbed, it effectively goes slower. But the light is not necessarily re-emitted in the same direction. Sometimes it is, but sometimes it is not. This creates a small, spherical wave (called a wavelet) that is emitted by that atom. So Huygens idea is not too bad.

We can use this idea for a single slit and look at what happens as the light goes through. Here is such a slit.

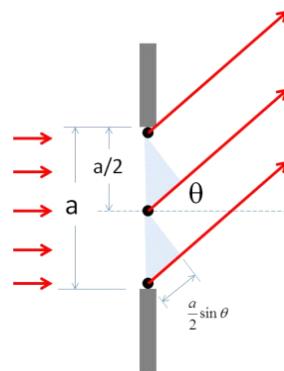
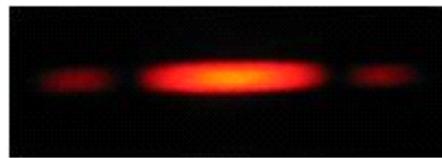


Figure 11.9.

In the figure above, we have divided a single slit of width a into two parts, each of size $a/2$. According to Huygens’ principle, each position of the slit acts as a source of light rays. So we can treat half a slit as two coherent sources. These two sources should

interfere. So what do we see when we perform such an experiment?



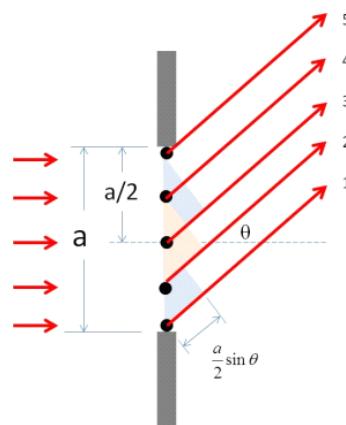
The figure shows a diffraction pattern for a thin slit. There are several terms that are in common use to describe the pattern

1. Central Maximum: The broad intense central band.
2. Secondary Maxima: The fainter bright bands to both sides of the central maxima
3. Minima: The dark bands between the maxima

Let's see how these structures are formed.

Narrow Slit Intensity Pattern

Let's use figure 11.9 to find the dark minima of the single slit pattern. First we should notice that figure 11.9 could have another set of rays that contribute to the bright spot because they will also have a path difference of $(a/2) \sin \theta$. Let's fill these in. They are rays 2 and 4 of the next figure.



Before we started with what we are now calling rays 1 and 3. Ray 1 travels a distance

$$\delta = \frac{a}{2} \sin(\theta) \quad (11.6)$$

farther than ray 3. As we just argued, rays 2 and 4 also have the same path difference, and so do rays 3 and 5. If this path difference is $\lambda/2$ then we will have destructive interference. The condition for a minima is then

$$\frac{a}{2} \sin(\theta) = \pm \frac{\lambda}{2} \quad (11.7)$$

or

$$\sin(\theta) = \pm \frac{\lambda}{a} \quad (11.8)$$

Now we could also divide the slit into four equal parts. Then we have a path difference of

$$\delta = \frac{a}{4} \sin(\theta) \quad (11.9)$$

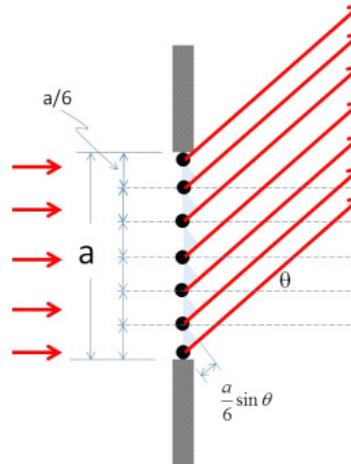
and to have destructive interference we need this path difference to be $\lambda/2$

$$\frac{a}{4} \sin(\theta) = \pm \frac{\lambda}{2} \quad (11.10)$$

or

$$\sin(\theta) = \pm \frac{2\lambda}{a} \quad (11.11)$$

We can keep going to find a minima at



$$\sin(\theta) = \pm \frac{3\lambda}{a} \quad (11.12)$$

and in general at

$$\sin(\theta) = m \frac{\lambda}{a} \quad m = \pm 1, \pm 2, \pm 3 \dots \quad (11.13)$$

Question 223.11.1

We only found the dark spots in a single slit intensity pattern. The bright spots must be in between the dark spots, but finding them is a little more trouble than finding the dark

spots. Do do this, let's lok at the intensity function for a single slit interference pattern.

Intensity of the single-slit pattern

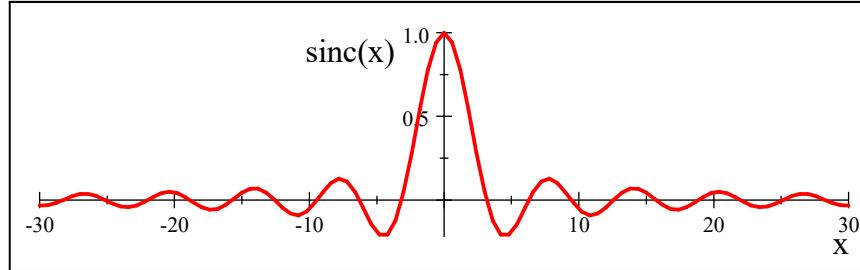
We could derive the single slit intensity pattern, it's not too hard to do. But instead I will just give the result here and we will interpret that result.

$$I = I_{\max} \left(\frac{\sin \left(\frac{\pi}{\lambda} a \sin \theta \right)}{\frac{\pi}{\lambda} a \sin \theta} \right)^2 \quad (11.14)$$

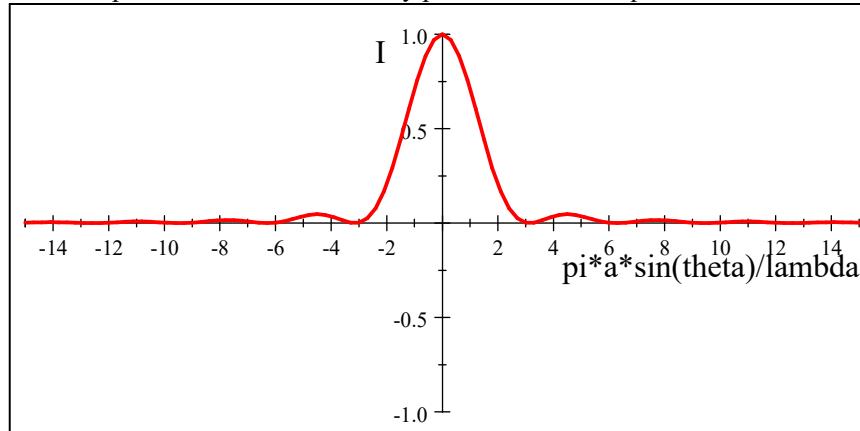
Notice this has the form

$$\frac{\sin x}{x}$$

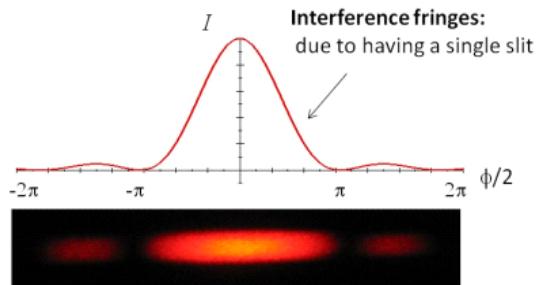
which has a distinctive shape.



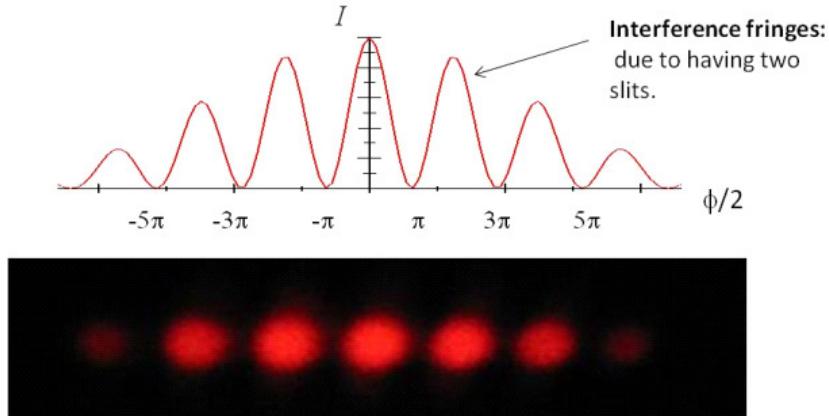
This is known as a sinc function (pronounced like “sink”). It has a central maximum as we would expect. Of course our intensity pattern has a sinc squared



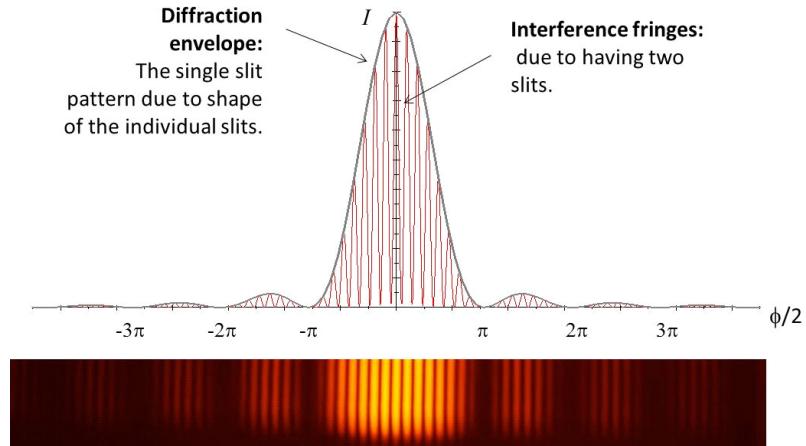
You can see the central maximum and the much weaker minima produced by this function. Indeed, it seems to match what we saw very well. Putting it all together, our pattern looks like this.



This is really an interesting result. You might wonder why, when we found the two slit interference pattern, there was no evidence of the single slit fringing that we discovered in this chapter. After all, a double slit system is made from single slits. Shouldn't there be some effect due to the fact that the slits are individually single slits? The answer is that we did see some hint of the single slit pattern. Remember the figure below.



The intensity of the peaks seems to fall off with distance from the center. We dealt with only the center-most part of the pattern. If we draw the pattern for larger angles, we see the following.



It takes a bright laser or dark room to see the secondary groups of fringes easily, but we can do it. We can also graph the intensity pattern. It is the combination of the two slit and single slit pattern with the single slit pattern acting as an envelope.

$$I = I_{\max} \cos^2 \left(\frac{\pi d \sin(\theta)}{\lambda} \right) \left(\frac{\sin \left(\frac{\pi a \sin(\theta)}{\lambda} \right)}{\frac{\pi a \sin(\theta)}{\lambda}} \right)^2 \quad (11.15)$$

Note that one of the double slit maxima is clobbered by a minimum in the single slit pattern. We can find out the order number of the missing maximum. Recall that

$$d \sin(\theta) = m\lambda$$

describes the maxima from the double slit. But

$$a \sin(\theta) = \lambda$$

describes the minimum from the single slit. Dividing these yields

$$\begin{aligned} \frac{d \sin(\theta)}{a \sin(\theta)} &= \frac{m\lambda}{\lambda} \\ \frac{d}{a} &= m \end{aligned}$$

so the

$$m = \frac{d}{a} \quad (11.16)$$

double slit maximum will be missing.

12 Apertures and Interferometers

Fundamental Concepts

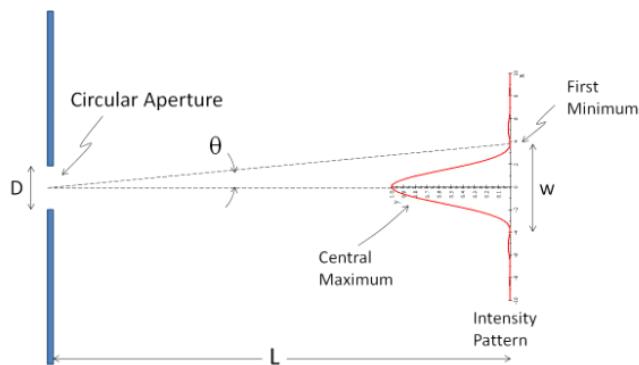
- Round apertures act very much like slits, with some added numerical factors
- If $\lambda \ll D$ we see little evidence for the wave nature of light. This limit is called the geometric optics limit or the ray approximation
- Interferometers can measure phenomenally small displacements using the wave nature of light

Circular Apertures

Question 223.12.1

Question 223.12.3

Our analysis of light going through holes has been somewhat limited by squarish holes or slits. But most optical systems, including our eyes don't have rectangular holes. So what happens when the hole is round? The situation is as shown in the next figure.



Before we discuss this situation, let's think about what we know about the width of a single slit pattern. We remember that

$$\sin(\theta) = (1) \frac{\lambda}{a}$$

for the first minima, or, because the angles are small,

$$\theta \approx \frac{\lambda}{a}$$

and from the figure we can see that

$$\tan \theta = \frac{y}{L}$$

or

$$\theta \approx \frac{y}{L}$$

so long as θ is small, then we find the position of the first minimum to be

$$y = \frac{\lambda}{a} L$$

This is the distance from the center bright spot to the first dark spot. The width of the bright spot is twice this distance

$$w = 2 \frac{\lambda}{a} L$$

We expect something like this for our circular aperture. The derivation for the circular aperture is not really too hard, but it involves Bessel functions, which are beyond the math requirement for this course. So I will give you the answer

$$\theta = 1.22 \frac{\lambda}{D}$$

where D is the diameter of the circular aperture (like a was the width of the slit) and as before

$$\tan \theta = \frac{y}{L}$$

so

$$\theta \approx \frac{y}{L}$$

which gives us a first minimum location of

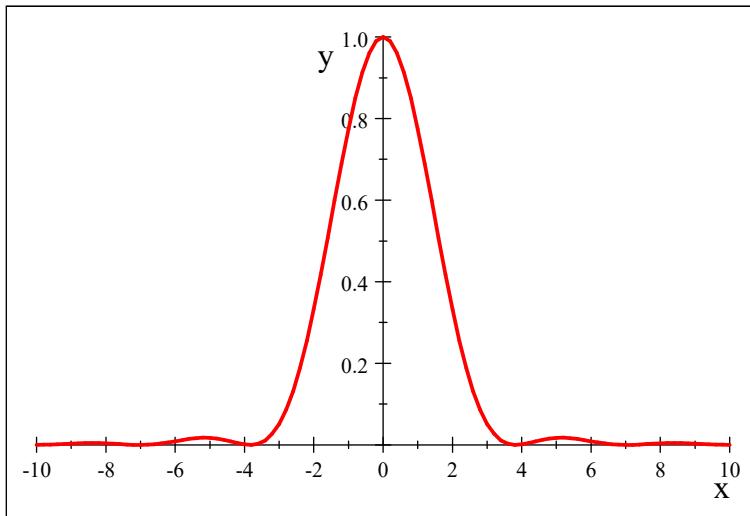
$$y = 1.22 \frac{\lambda}{D} L \quad (12.1)$$

and a width of

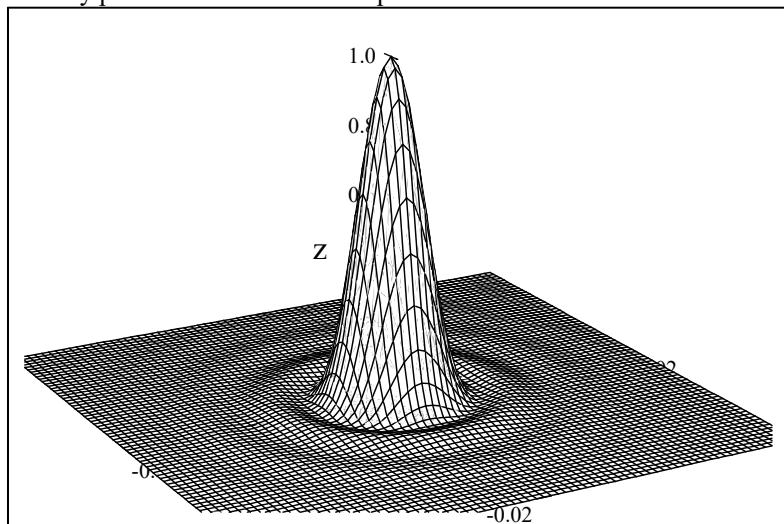
$$w = 2.44 \frac{\lambda}{D} L \quad (12.2)$$

Airy Pattern Demo

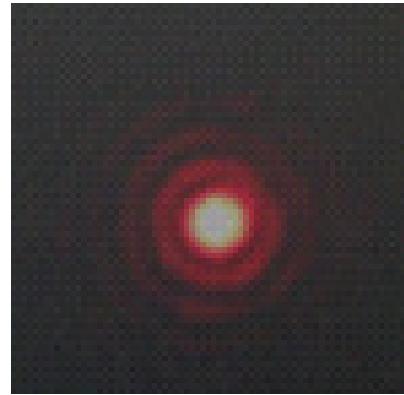
The picture in most books is a little bit deceptive. The pattern looks a little like the slit pattern. But the secondary maxima are actually very small for the circular aperture case. Much smaller than the secondary maxima in the slit case. Here is a larger version of a cross section of the intensity pattern.



Notice how small the secondary and tertiary maxima are. A three dimensional version of the intensity pattern from the circular aperture.



With a bright enough laser, they pattern becomes visible.



Interferometers

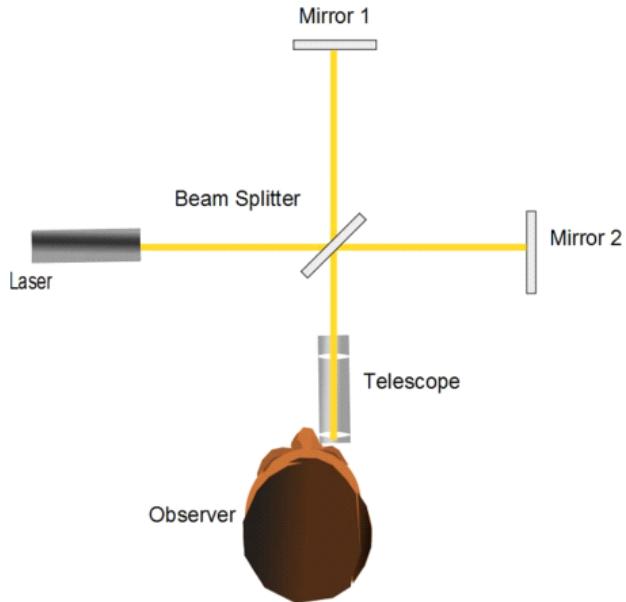
Interferometer

Demo

Before we leave wave properties of physics and go to the ray approximation, we should study some devices that use interference.

The Michelson Interferometer

The Michelson interferometer is another device that uses path differences to create interference fringes.



The device is shown in the figure. A coherent light source is used. The light beam is split into two beams that are usually at 90° apart. The beams are reflected off of two mirrors back along the same path and are mixed at the telescope. The result (with perfect alignment) is a target fringe pattern like the first two shown below.



If the alignment is off, you get smaller fringes, but the system can still work. This is shown in the last image in the previous figure.

In the figure, we have constructive interference in the center, but if we move one of the mirrors half a wavelength, we would have destructive interference and would see a dark spot in the center. This device gives us the ability to measure distances on the order of the wavelength of the light. When the distance is continuously changed, the pattern seems to grow from the center (or collapse into the center).

Notice that if the mirror is moved $\frac{\lambda}{2}$, the path distance changes by λ because the light

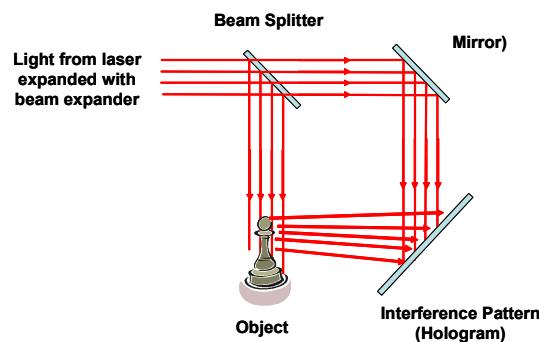
travels the distance to the mirror and then back from the mirror (it travels the path twice!).

Holography

Hologram demo-
picture of woman

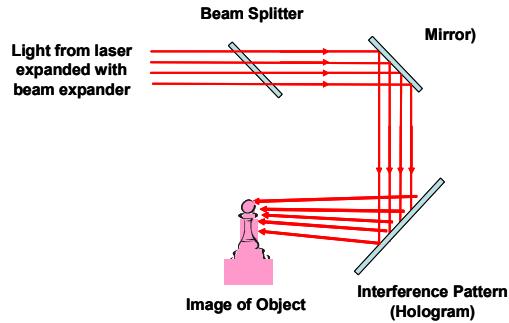
Hologram demo-
chess pieces

You may have seen holograms in the past. We have enough understanding of light to understand how they are generated now.



A device for generating a hologram is shown in the figure above. Light from a laser or other coherent source is expanded and split into two beams. One travels to a photographic plate, the other is directed to an object. At the object, light is scattered and the scattered light also reaches the photographic plate. The combination of the direct and scattered beams generates a complicated interference pattern.

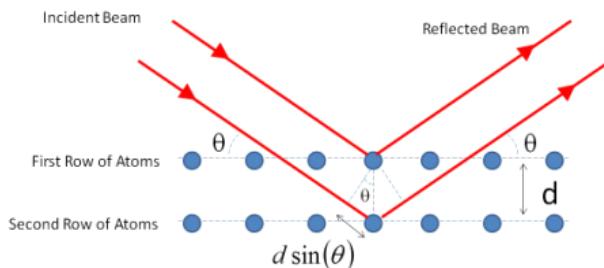
The pattern can be developed (like you develop photographic film). Once developed, it can be re-illuminated with a direct beam. The emulsion on the plate creates complicated patterns of light transmission, which combine to create interference. It is like a very complicated slit pattern or grating pattern. The result is a three-dimensional image generated by the interference. The interference pattern generates an image that looks like the original object.



Diffraction of X-rays by Crystals

If we make the wavelength of light very small, then we can deal with very small diffraction gratings. This concept is used to investigate the structure of crystals with x-rays. The crystal lattice of molecules or atoms creates the regular pattern we need for a grating. The pattern is three dimensional, so the patterns are complex.

Let's start with a simple crystal with a square regular lattice. $NaCl$ has such a structure.



If we illuminate the crystal with x-rays, the x-rays can reflect off the top layer of atoms, or off the second layer of atoms (or off any other layer, but for now let's just consider two layers). If the spacing between the layers is d , then the path difference will be

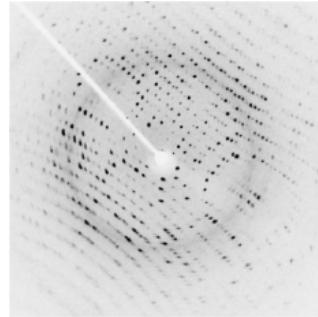
$$\delta = 2(d \sin(\theta)) \quad (12.3)$$

then for constructive interference

$$2d \sin(\theta) = m\lambda \quad m = 1, 2, 3, \dots \quad (12.4)$$

This is known as *Bragg's law*. This relationship can be used to measure the distance between the crystal planes.

A resulting pattern is given in the following figure.



Diffraction image of protein crystal. Hen egg lysozyme, X-ray source Bruker I μ S,
 $\lambda = 0.154188$ nm, 45 kV, Exposure 10 s.
DNA makes an interesting diffraction pattern.

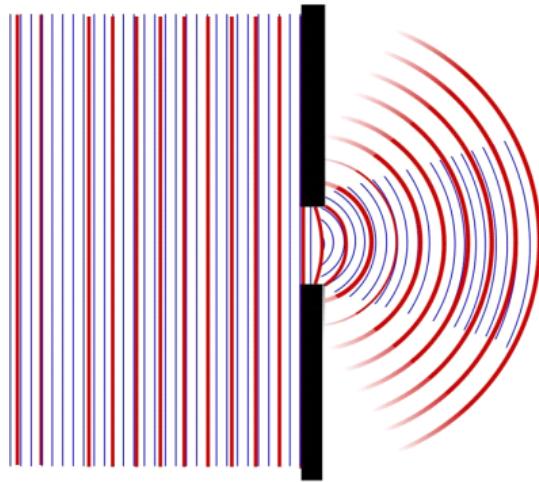


X-ray diffraction pattern of DNA

Transition to the ray model

Question 223.12.4

In the next figure, two waves of different wavelets go through a single opening. The wave representing the central maximum is shown in each case, but not the secondary maxima.



Notice that the smaller wavelength has a narrower central maxima as we would expect from

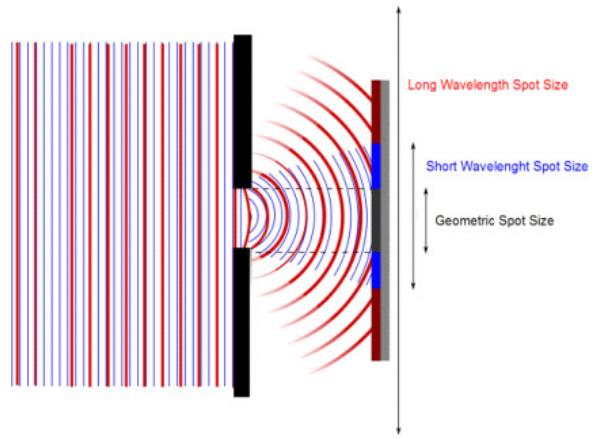
$$\sin(\theta) = 1.22 \frac{\lambda}{D}$$

or

$$\theta \approx 1.22 \frac{\lambda}{D}$$

we see that the ratio of the wavelength to the hole size determines the angular extent of the central maxima. The smaller the ratio, the smaller the central region. We can use this to explain why the wave nature of light was so hard to find.

The patch of light on a screen that is created by light passing through the aperture is created by the central maximum.



For the long wavelength (red) the central maximum is larger than the screen. The short wavelength spot will be wholly on the screen as shown. The geometric spot is what

we would see if the light traveled straight through the opening. Notice that the short wavelength spot is closer to the size of the geometric spot. In the limit that

$$\lambda \ll a$$

or for circular openings

$$\lambda \ll D$$

then

$$\theta \approx \frac{\lambda}{a} \approx 0$$

or

$$\theta \approx \frac{\lambda}{D} \approx 0$$

and the spot size would be very nearly equal to the geometric spot size.

This is the limit we will call the *ray approximation*.

For most of mankind's time on the Earth, it was very hard to build holes that were comparable to the size of a wavelength of visible light. So it is no wonder that the waviness of light was missed for so many years.

But this ray limit is very useful for apertures the size of camera lenses. So starting next lecture we will begin to use this small λ , large aperture approximation.

13 Ray Model

Fundamental Concepts

- When the aperture size is given by $D_{aperture} = \sqrt{2.44\lambda L}$ we are at a critical size bounding the geometric and wave optics regions
- Coherent light is light that maintains a common phase, direction, and wavelength.
- Light reflects from a specular surface with equal angles

The Ray Approximation in Geometric Optics

Last time we said that when the geometric spot size was larger than the spot due to diffraction, we could ignore diffraction and use the simpler ray model. This is usually true in our personal experiences. But this may not be true in experiments or devices we design. We should see where the crossover point is.

Intuitively, if the aperture and the spot are the same size, that ought to be some sort of critical point. That is when the aperture size is equal to the spot size

$$D_{aperture} = 2.44 \frac{\lambda}{D_{aperture}} L$$

This gives

$$D_{aperture} = \sqrt{2.44\lambda L}$$

Of course this is for round apertures, but for square apertures we know we remove the 2.44. This gives about a millimeter for visible wavelengths.

$$\begin{aligned} D_{aperture} &= \sqrt{2.44(500 \text{ nm})(1 \text{ m})} \\ &= 1.1045 \times 10^{-3} \text{ m} \end{aligned}$$

for apertures much larger than a millimeter, we expect interference effects due to diffraction through the aperture to be much harder to see. We expect them to be easy to see if the aperture is smaller than a millimeter. But what about when the aperture is

about a millimeter in size? That is a subject for PH375, and so we will avoid this case in this class. But this is not too restrictive. Most good optical systems have apertures larger than 1 mm. Cell phone cameras may be an exception (but I don't consider cell phone cameras to be good optical systems). Even our eyes have an aperture that varies from about 2 mm to about 7 mm, so most common experiences in visible wavelengths will work fine with what we learn. Note that for microwave or radio wave systems this may really not be true!

How about the other extreme? Suppose $\lambda \gg D$. This is really beyond our class (requires partial differential equations), but in the extreme case, we can use reason to find out what happens. If the opening is much smaller than the wavelength, then the wave does not see the opening, and no wave is produced on the other side. This is the case of a microwave oven door. If the wavelength is much larger than the spacing of the little dots or lines that span the door, then the waves will not leave the interior of the microwave oven. Of course as the wavelength becomes closer to D this is less true, and this case is more challenging to calculate, and we will save it for a 300 level electrodynamics course.

To summarize

- $\lambda \ll D$ Wave nature of light is not visible
- $\lambda \approx D$ Wave nature of light is apparent
- $\lambda \gg D$ Little to no penetration of aperture by the wave

We can see that early researchers might not have spent a lot of time with sub-millimeter sized holes, so the wave nature of light was not as apparent to Newton and his contemporaries.

The ray model and phase

Question 223.13.1

There is a further complication that helps to explain why the wave nature of light was not immediately apparent to early researchers. Let's consider a light source.



For a typical light source, the filament or light emitting diode (LED) is larger than about a millimeter, which is much larger than the wavelength. So, we should already expect that diffraction might be hard to see. But the filament is made of hot metal (we will

leave LED workings for another class). The atoms of the hot metal emit light because of the extra energy they have. The method of producing this light is that the atom's excited electrons are in upper shells because of the extra thermal energy provided by the electricity flowing through the filament. But the electrons eventually fall to their proper shell, and in doing so they give off the extra energy as light. It is not too hard to believe that this process of exciting electrons and having them fall back down is a random process. Each electron that moves starts a wave. The atoms have different positions, so there will be a path difference Δr between each atom's waves. There will also be a time difference Δt between when the waves start. We can model this with a $\Delta\phi_o$.

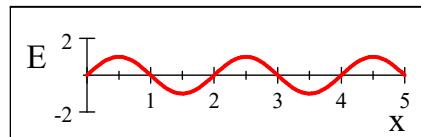
Question 223.13.2

It is also true that not all of the electrons fall from the same shell. This gives us different frequencies, so we expect beating between different waves from different atoms. It is also true that we have millions of atoms, so we have millions of waves.

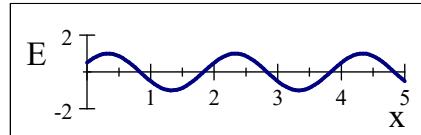
Let's look at just two of these waves

$$\begin{aligned}\lambda &= 2 \\ k &= \frac{2\pi}{\lambda} \\ \omega &= 1 \\ \phi_o &= \frac{\pi}{6} \\ t &= 0 \\ E_o &= 1 \frac{N}{C}\end{aligned}$$

$$E_1 = E_{\max} \sin(kx - \omega t)$$



$$E_2 = E_{\max} \sin(kx - \omega t + \phi_o)$$

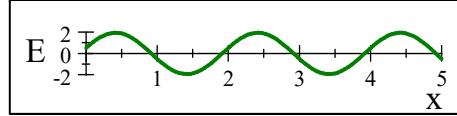


then

$$E_r = E_{\max} \sin(kx - \omega t) + E_o \sin(kx - \omega t + \phi_o)$$

We found a nice meaningful way to write the resultant wave.

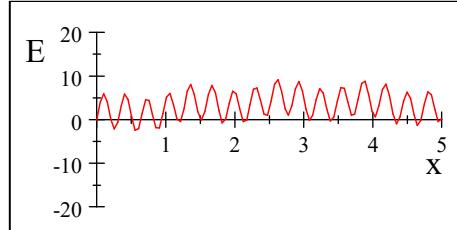
$$E_r = 2E_{\max} \cos\left(\frac{\phi_o}{2}\right) \sin\left(kx - \omega t + \frac{\phi_o}{2}\right)$$



But suppose we complicate the situation by sending out lots of waves at random times, each with different amplitudes and wavelengths. If we look at a single point for a specific time, we might be experiencing interference, but it would be hard to tell. Lets try this mathematically. I will combine many waves with random phases, some coming from the right and some coming from the left.

$$\begin{aligned} E_1 = & E_{\max} \sin\left(5x - \omega t + \frac{\pi}{4}\right) + 0.5E_{\max} \sin\left(0.2x - \omega t - \frac{\pi}{6}\right) \\ & + 3.6E_{\max} \sin\left(0.4x - \omega t + \frac{\pi}{10}\right) + 4E_{\max} \sin\left(20x - \omega t - \frac{\pi}{7}\right) \\ & + 0.2E_{\max} \sin\left(15x - \omega t + 1\right) + 0.7E_{\max} \sin\left(0.7x - \omega t - 0.25\right) \end{aligned}$$

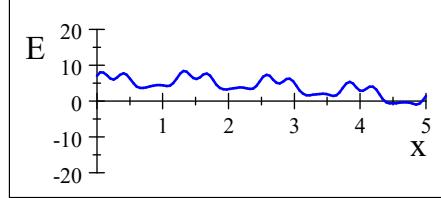
Here is what E_1 would look like.



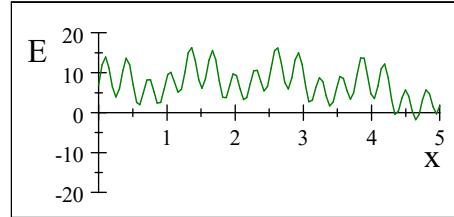
And now let's make another random wave, E_2

$$\begin{aligned} E_2 = & E_{\max} \sin\left(0.2x + \omega t + \pi\right) + 2E_{\max} \sin\left(5x + \omega t + \frac{\pi}{6}\right) \\ & + 6E_{\max} \sin\left(0.4x + \omega t + \frac{\pi}{3.5}\right) + 0.4E_{\max} \sin\left(20x + \omega t - 0\right) \\ & + E_{\max} \sin\left(15x + \omega t + 1\right) + 0.7E_{\max} \sin\left(0.7x + \omega t - 4\right) \end{aligned}$$

Which looks like this



Then $E_1 + E_2$ looks like



In this example, you could think about the superposition of E_1 and E_2 and predict the outcome, but if there were millions of waves, each with its own wavelength, phase, and amplitude, the situation would be hopeless. Note that the fluctuations in these waves are much more frequent than our original waves. With all the added waves, we get a rapid change in amplitude.

Now if these waves are light waves, our eyes and most detectors are not able to react fast enough to detect the rapid fluctuations. So if there is constructive or destructive interference that might be simple enough to distinguish, the interference pattern will change so fast that we will miss it due to our detection systems' integration times. To describe this rapidly fluctuating interference pattern that we can't track with our detectors, we just say that light bulbs emit *incoherent light*. The ray approximation assumes incoherent light.

But then light bulbs and hot ovens and most things must emit incoherent light. Does any thing emit coherent light? Sure, today the easiest source of coherent light is a laser. That is why I have used lasers in the class demonstrations so far. Really though, even a laser is not perfectly coherent. One property of the laser is that it produces light with a long *coherence length*, or it produces light that can be treated under most circumstances as being monochromatic and having a single phase across the wave for much of the beam length. Radar and microwave transmitters emit coherent light (but at frequencies we can't see) and so do radio stations.

In the past, one could carefully create a monochromatic beam with filters. Then split the beam into two beams and remix the two beams. This would generate two mostly coherent sources if the distances traveled were not too large. This is what Young did.

Coherency

Question 223.13.3

To be coherent,

1. A given part of the wave must maintain a constant phase with respect to the rest of the wave.
2. The wave must be monochromatic

These are very hard criteria to achieve. Most light, like that from our light bulb, is not coherent.

Reflection

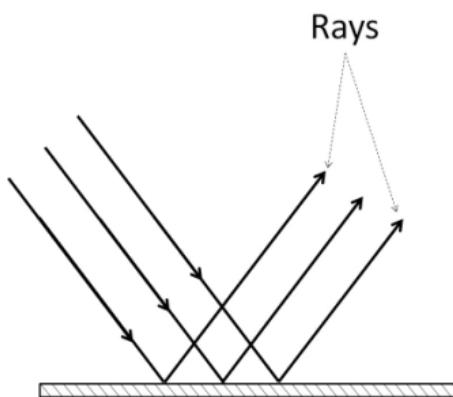
Question 223.13.4

Question 223.13.5

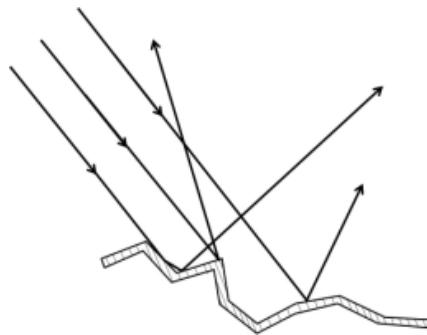
Specular and Diffuse Reflector Demo

In the *Star Wars* movies inter-galactic star ships blast each other with laser cannons. The laser beams streak across the screen. This is dramatic, but not realistic. For us to see the light, some of the light must get to our eyes. The light must either travel directly to our eyes from the source, or it must bounce off of something.

Using the ray approximation we wish to find what happens when a bundle of rays reaches a boundary between media. If the media boundary is very smooth, then the rays are reflected in a uniform way. This is called *specular reflection*



If it is not smooth, then something different happens. The rays are reflected, but they are reflected randomly



Question 223.13.6

This is called *diffuse reflection*

This difference can be seen in real life



We said the surface must be smooth for there to be specular reflection. What does smooth mean? Generally the size of the rough spots must be much smaller than a wavelength to be considered smooth. So suppose we have a red laser. How small do the surface variations have to be for the surface to be considered smooth? The wavelength of a *HeNe* laser is

$$\lambda_{HeNe} = 633 \text{ nm}$$

This is very small. Modern optics for remote sensing are often manufactured to 1/10 of a wavelength, which would be 63 nm.

How about a microwave beam of light like your cell phone uses?

$$\begin{aligned} c &= \lambda f \\ \lambda &= \frac{c}{f} = \frac{3 \times 10^8 \frac{\text{m}}{\text{s}}}{1 \text{ GHz}} \\ &= 0.3 \text{ m} \end{aligned}$$

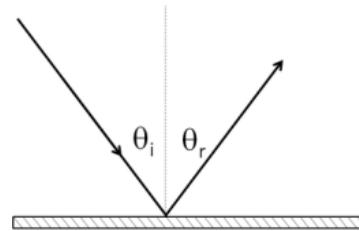
We can see that we must be careful in our definition of “smooth.”

Law of reflection

Ball Bounce Demo

Experience shows that if we do have a smooth surface, that light bounces much like a ball. This is why Newton thought light was a particle. Suppose we take a flat surface and we shine a light on it. We have a ray that approaches at an angle θ_i measured from the normal. Then the reflected ray will leave the surface with an angle θ_r measured from the normal such that

$$\theta_r = \theta_i$$



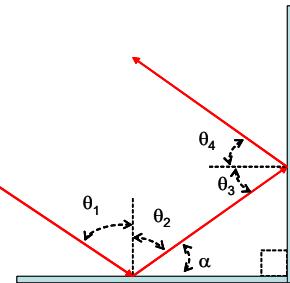
Question 223.13.7

This is called the *law of reflection*.

Question 223.13.8

Retroreflection

Let's take an example



Let's take our system to be two mirrors set at a right angle. We have a beam of light incident at angle θ_1 . By the law of reflection, it must leave the mirror at $\theta_2 = \theta_1$. We can see that α must be $90^\circ - \theta_2$ and it is clear that $\theta_3 = \alpha$. By the law of reflection, $\theta_3 = \theta_4$. Then, since

$$\begin{aligned} 90^\circ &= \theta_2 + \alpha \\ &= \theta_2 + \theta_3 \end{aligned}$$

and

$$90^\circ = \theta_1 + \theta_4$$

then the total angular change is

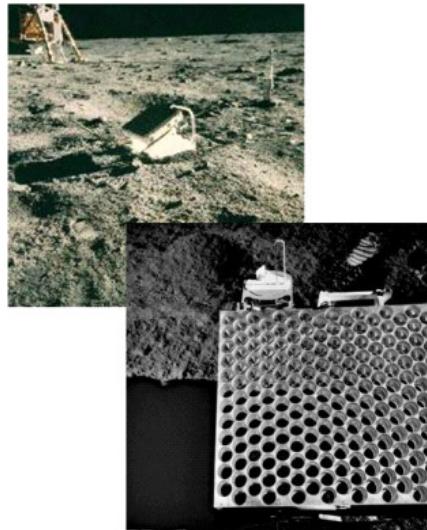
$$90^\circ + 90^\circ = 180^\circ$$

or the outgoing ray is sent back toward the source! If we do this in three dimensions we have a corner cube.



Radar retroreflector tower located in the center of Yucca Flat dry lake bed. Used as a radar target by maneuvering aircraft during "inert" contact fusing bomb drops at Yucca Flat. Sandia National Laboratories conducted the tests on the lake bed from 1954 to 1956. (Image in the Public Domain in the United States)

The figure above is a radar corner cube set. The one below is an optical corner cube set on the moon.



Apollo Retroreflector (Images in the Public Domain courtesy NASA)

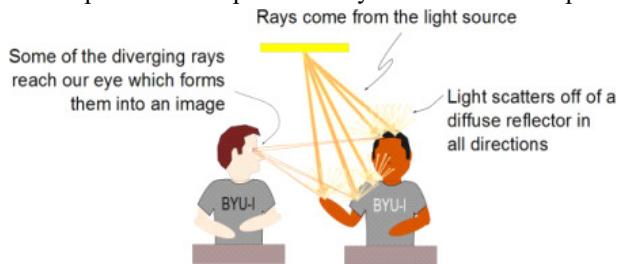
We use this optical corner cube array to reflect light off of the moon. The time it takes the light to go to the moon and back can be converted into an Earth-Moon distance for monitoring how close the moon is to the Earth.

Question 223.13.9

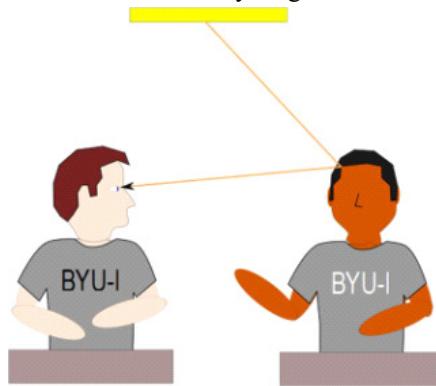
Reflections, Objects, and seeing

Armed with the law of reflection, we can start to understand how we see things. Using the ray concept, we can say that a ray of light must leave the light source. That ray then reflects from something. Suppose you look at the person sitting next to you in class. Light from the ceiling lights has reflected from that person. But is the person a specular or diffuse reflector?

Once again, we can only give an answer relative to the wavelength of light. For visible light, your neighbors do not look like mirrors. They are diffuse reflectors. Light bounces off of them in every direction. Your eye is designed to take this diverging set of rays and condense it into a picture of the person that your brain can interpret.



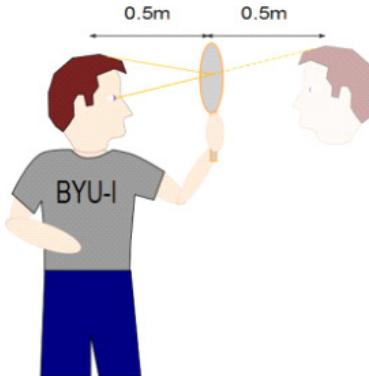
We tend to not draw the rays that bounce off the diffuse reflector but that don't get to our eyes, because we don't see them. So a ray diagram is usually much simpler.



This is easy to understand, but we must keep in mind the wildly fluctuating waviness that is masked by our macroscopic view.

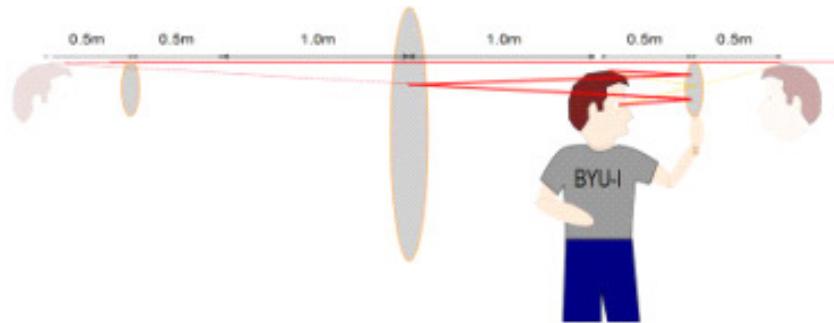
Question 223.13.10

We can use the idea of a ray diagram to solve problems. Suppose you hold a mirror half a meter in front of you and look at your reflection. Where would the reflection appear to be?



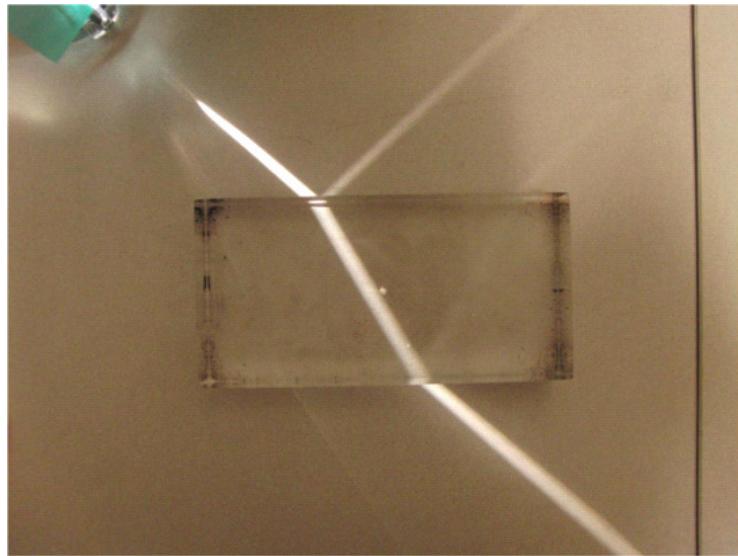
Question 223.13.11

Knowing that rays travel in straight lines and that our mind interprets rays as going in straight lines, then we can use rays to see where the light appears to be from. The image is half a meter behind the mirror. Now suppose we look at an image of that image in a mirror behind us.



The ray diagram makes it easy to see that the image will appear to be 2 m behind the big mirror.

14 Refraction and images



We studied light reflecting from a surface. We can see the reflection in the image above. But light also is transmitted through the piece of glass in the figure. Note the change in direction at the interfaces. This is penetration of a material by light is called refraction, and will be the subject of this lecture.

Fundamental Concepts

- Refraction is a change of direction of a light ray as it crosses an interface
- The wavelength of the light changes at an interface
- The angle changes according to Snell's law $n_i \sin \theta_i = n_t \sin \theta_t$
- When going from a high index to a low index material, the light may totally reflect, with no transmission
- Refraction can form images

Refraction

Not all surfaces reflect all the light. Some, like the lenses shown below, reflect some light at visible wavelengths, but are transparent so most of the light travels through them.

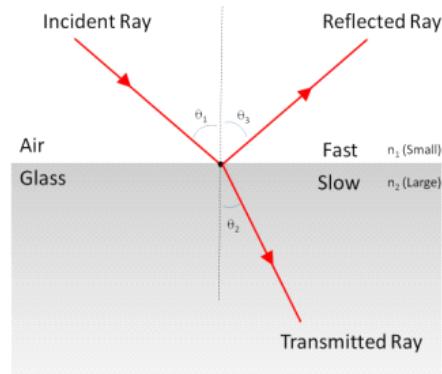


We need a way to deal with transparent materials. This is tricky, because different wavelengths of light penetrate different materials in different ways. As an example, this is also a lens



IR lens. (Image in the Public Domain, courtesy US Navy)
but it clearly is not transparent at visible wavelengths. But it is transparent in the infrared. So what might be transparent at one wavelength might not be at another.

When light travels into a material, we say it is transmitted. The situation is shown schematically below.



In the figure we see a ray incident on an air-glass boundary. Some of the light is reflected just as we saw before. But some passes into the glass. Notice that the angle between the normal and the new transmitted ray is *not* equal to the incident ray. We say the ray has been bent or *refracted* by the change of media. Many experiments were performed to find a relationship between the incident and the refracted angles. It was found that

$$\frac{\sin(\theta_2)}{\sin(\theta_1)} = \frac{v_2}{v_1} = \text{constant} \quad (14.1)$$

Many optics books write this as

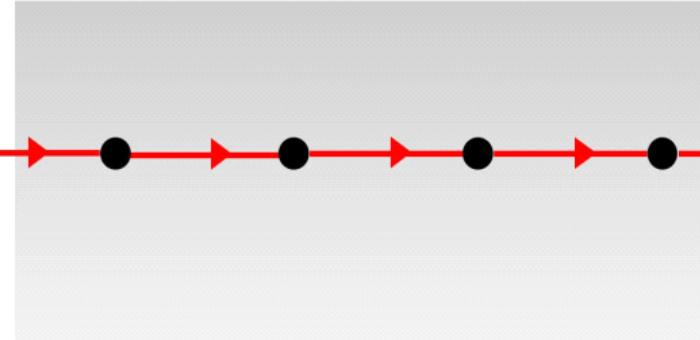
$$\frac{\sin(\theta_t)}{\sin(\theta_i)} = \frac{v_2}{v_1} = \text{constant} \quad (14.2)$$

where the subscript *i* stands for “incident” and the subscript *t* stands for “transmitted.” Note that we are using the fact that the average speed of light changes in a material. We should probably recall why this should occur

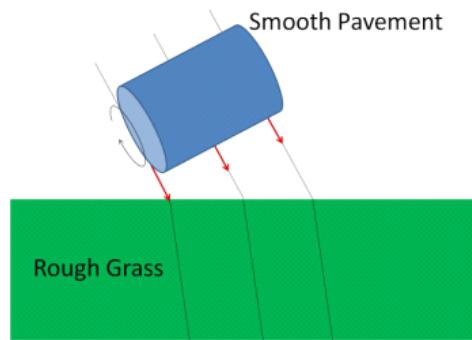
Speed of light in a material

In a vacuum, light travels as a disturbance in the electromagnetic field with nothing to encounter. In a material (like glass) the light waves continually hit atoms. We have not studied antennas, but I think many of you know that an antenna works because the electrons in the metal act like driven harmonic oscillators. The incoming radio waves drive the electron motion. Here each atom has electrons, and the atoms act like little antennas, their electrons moving and absorbing the light. But the atom cannot keep the extra energy (PH433), so it is readmitted. It travels to the next atom and the process repeats. Quantum mechanics tells us that there is a time delay in the re-emission of the light. This causes a secondary wave to mix with the incoming wave. The combined result is that the propagation energy in the light wave slows down. Thus the speed of light is

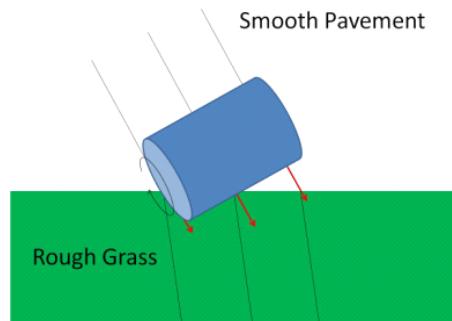
slower in a material.



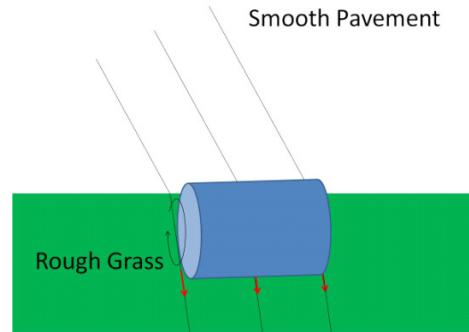
But why does this slowing cause the light ray to bend? As a mechanical analog, consider a rolling barrel.



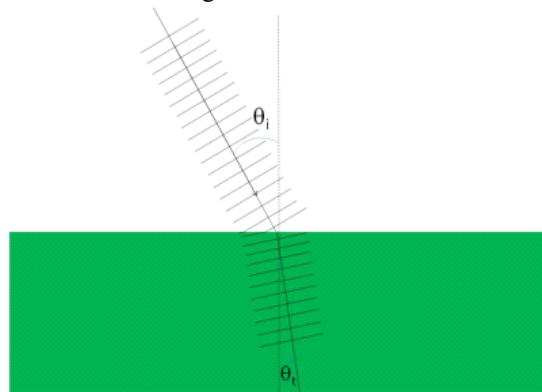
As the barrel rolls from a flat low-friction concrete to a higher-friction grass lawn, the friction slows the barrel. If the barrel hits the lawn parallel to the boundary (so its velocity vector is perpendicular to the boundary), then the barrel continues in the same direction at the slower speed. But if it hits at an angle, the leading edge is slowed first.



This makes the trailing edge travel faster than the leading edge, and the barrel turns slightly.



We expect the same behavior from light.



We can see that the left hand side of the wave hits the slower (green) material first and slows down. The rest of the wave front moves quicker. The result is the turning of the wave.

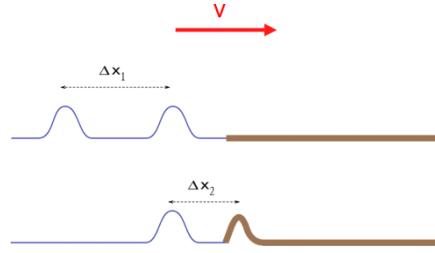
Question 223.14.1

Change of wavelength

We have found that when a wave enters a material, its speed may change. But we remember from wave theory

$$v = \lambda f \quad (14.3)$$

But it is time to review: does λ change, or does f change? If you will recall, we found that the change in speed at the boundary changes the wavelength. Recall that if we go from a fast material to a slow material, the forward part of the wave slows and the rest of the wave catches up to it.



This will compresses pulses, and lower the wavelength. Now that we know more about light we can also argue that f cannot change because

$$E = hf$$

If f changed, then we would either require an input of energy or we would store energy at the boundary because

$$\Delta f = \frac{\Delta E}{h}$$

This can't be true. If the wavelength changes, there is no such change in energy.

Since

$$v_1 = \lambda_1 f$$

and

$$v_2 = \lambda_2 f$$

then the ratio

$$\frac{v_1}{v_2} = \frac{\lambda_1}{\lambda_2}$$

and we again have our solution for the wavelength in the material

$$\lambda_2 = \lambda_1 \frac{v_2}{v_1}$$

which agrees with our previous analysis.

Index of refraction and Snell's Law

Question 223.14.2

Question 223.14.3

Because the equation

$$\frac{\sin(\theta_2)}{\sin(\theta_1)} = \frac{v_2}{v_1} = \text{constant}$$

has a constant ratio of velocities, it is convenient to define a term that represents that ratio. We already have a concept that can help. The *index of refraction* is just such a term. It assumes that one speed is the speed of light in vacuum, c .

$$n \equiv \frac{c}{v}$$

Then for our example

$$\frac{\sin(\theta_2)}{\sin(\theta_1)} = \frac{cv_2}{cv_1} = \frac{n_1}{n_2}$$

or

$$\frac{\sin(\theta_2)}{\sin(\theta_1)} = \frac{1}{n_2}$$

Suppose we don't have a vacuum (or air that is close to a vacuum). We can write our formula as

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \quad (14.4)$$

where we have determined

$$n_1 = \frac{c}{v_1}$$

and

$$n_2 = \frac{c}{v_2}$$

This is called *Snell's law of refraction* after the scientist who experimentally determined the relationship.

Again let's consider our wavelength change. Using the index of refraction we can write our equation relating the ratio of velocities and wavelengths as

$$\frac{v_1}{v_2} = \frac{\lambda_1}{\lambda_2} = \frac{\frac{c}{n_1}}{\frac{c}{n_2}} = \frac{n_2}{n_1}$$

which gives

$$\lambda_1 n_1 = \lambda_2 n_2$$

and if we have vacuum and a single material we can find the index of refraction from

$$n = \frac{\lambda}{\lambda_{material}} \quad (14.5)$$

Question 223.14.4

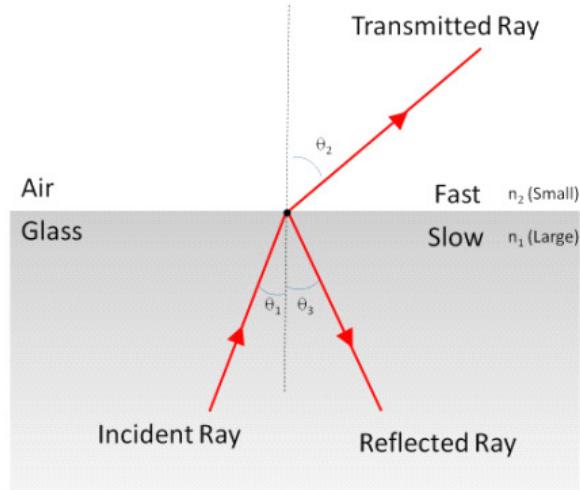
where $\lambda_{material}$ is the wavelength in the material.

Question 223.14.5

Total Internal Reflection

Question 223.14.6

Up to now we have assumed that light was coming from a region of low index of refraction into a region of high index of refraction. We should pause to look at what can happen if we go the other way.



We start with Snell's law

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

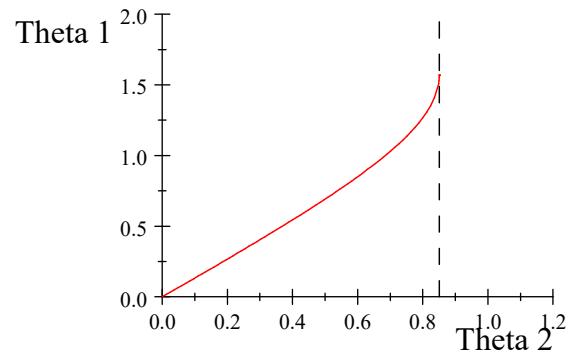
but this time $n = n_1$ and $n_2 \approx 1$ so

$$n \sin \theta_1 = \sin \theta_2$$

which gives

$$\theta_2 = \sin^{-1} (n \sin \theta_1) \quad (14.6)$$

If we take $n = 1.33$ (water) we can plot this expression as a function of θ_1



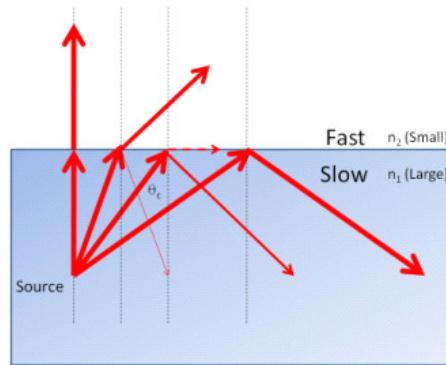
we see that at $\theta_1 = 0.85091 \text{ rad}$ (48.754°) the curve becomes infinitely steep. If we

use this value in our equation this gives

$$\begin{aligned}\theta_2 &= \sin^{-1}(n \sin(0.85091)) \\ &= 1.5708 \text{ rad} \\ &= 90^\circ\end{aligned}\quad (14.7)$$

Internal Reflection
Demo

The light skims along the edge of the water!



We can find the value of θ_1 that makes this happen without graphing. Set $\theta_2 = 90^\circ$ then

$n_1 \sin \theta_1 = n_2 \sin \theta_2$
becomes

$$\begin{aligned}n_1 \sin \theta_1 &= \sin(90^\circ) = 1 \\ \sin \theta_1 &= \frac{1}{n_1}\end{aligned}$$

so then θ_1 is given by

$$\theta_1 = \theta_c \equiv \sin^{-1}\left(\frac{1}{n_1}\right) \quad (14.8)$$

We give this value of θ_1 a special name. It is the *critical angle* for internal reflection. But what happens if we go farther than this ($\theta_1 > \theta_c$)? We will no longer have a transmitted ray. The ray will be reflected. This is why when you dive into a pool and look up, you see a region of the roof of the pool area (or sky) but off to the side of the pool the surface looks mirrored. It is also why you sometimes see the sides of a fish tank appear to be mirrored when you look through the front.



It is also why cut gems (like diamonds) sparkle. They capture the light with facets that are cut at angles that create total internal reflection. The light that enters the gem comes back out the front (We will study how to make the pretty colored sparkles next time).

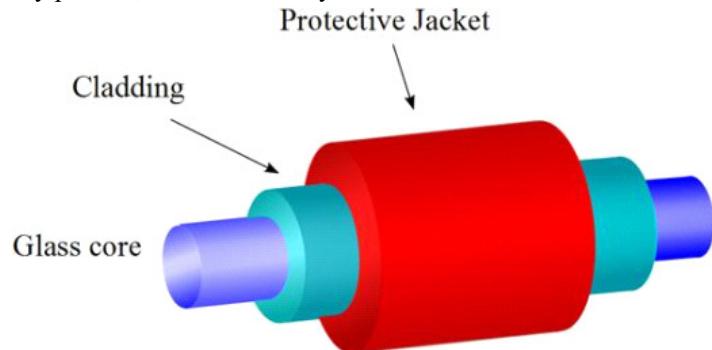
Question 223.14.7

Question 223.14.8

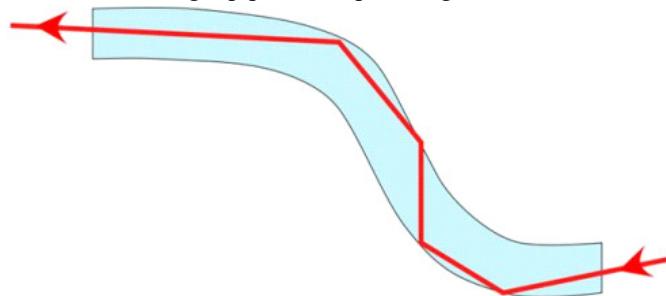
Question 223.14.9

Fiber Optics

Beyond pretty pebbles, this effect is very useful! It is the heart and soul of fiber optics.



An interior material with a lower index of refraction is inclosed in a cladding with a higher index. This creates a light pipe that traps the light in the fiber.



Giant Fiber Demo

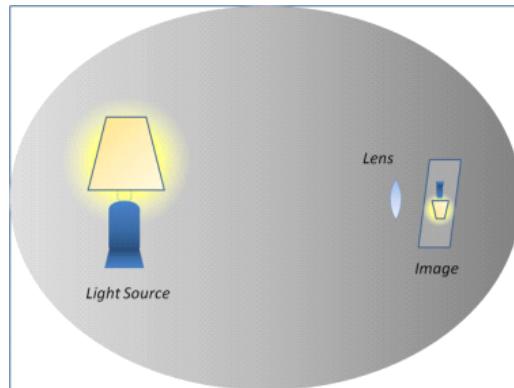
Modern fibers don't always have a hard boundary. The fibers have a gradual change in index of refraction that changes the direction of the light gradually. This keeps the light in the fiber but tends to direct along the fiber so the beam is not crisscrossing as it goes.

The cutting edge of fiber design today uses hollow fibers or fibers filled with different index material.

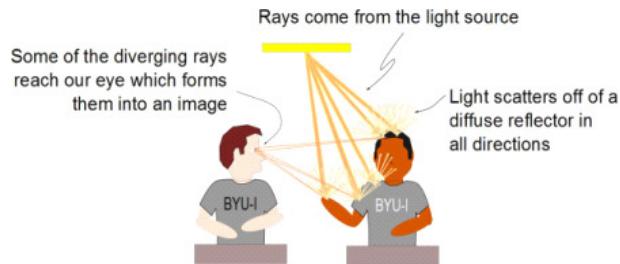
Images Formed by Refraction

Make Images with
Lens Demo

Let's think about what an image is. Take a piece of paper and a lens, and hold up the lens in a darkened room that has some bright object in it. Move the lens or the paper back and forth, and at just the right distance, a miniature picture of the bright object will appear. We should think about what the word "picture" means in this sense.



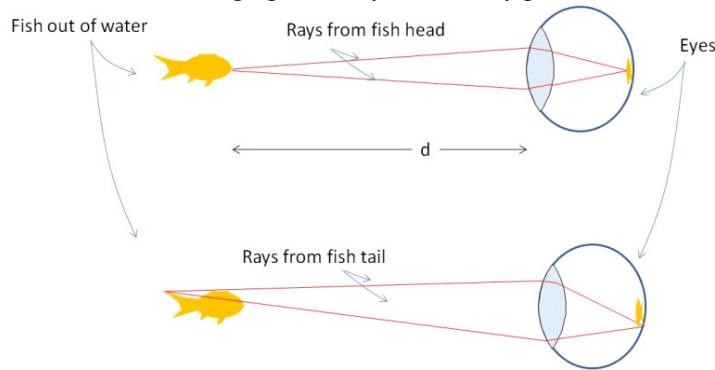
We have talked about how we see something. Remember the BYU-I guys from last time.



Our eyes gather rays that are diverging from the object because light has bounced off of the object. Our eyes intersect a diverging set of rays that form a definite pattern. That diverging set of rays forming a pattern is the picture of the object.

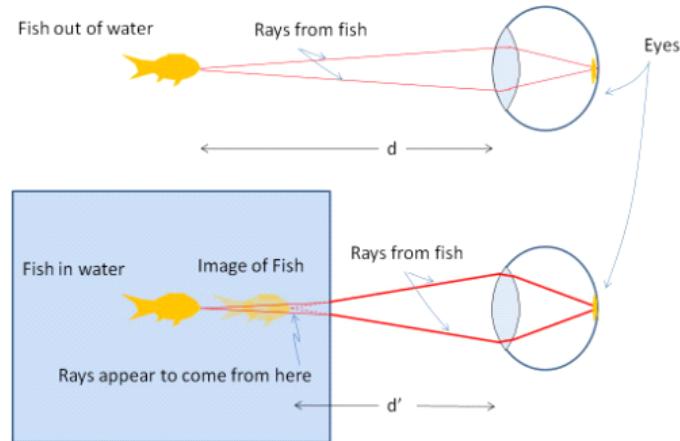
So when we say that the lens has formed a miniature picture of our dark room object, we mean that the lens has somehow formed a diverging set of rays that form a pattern that looks like the pattern formed by the diverging set of rays coming from the object, itself. In other words, the object forms a diverging set of rays, as normal, and our lens forms a duplicate set of rays in the same pattern, so we see the same thing. The lens' version is smaller, and upside down, but it is still essentially the same pattern.

As a first step to see how this works, consider our fish tank again. It would be bad on the fish, but think about looking at a fish in air. The room light would bounce off of the fish, and we would have a diverging set of rays from every point on the fish.



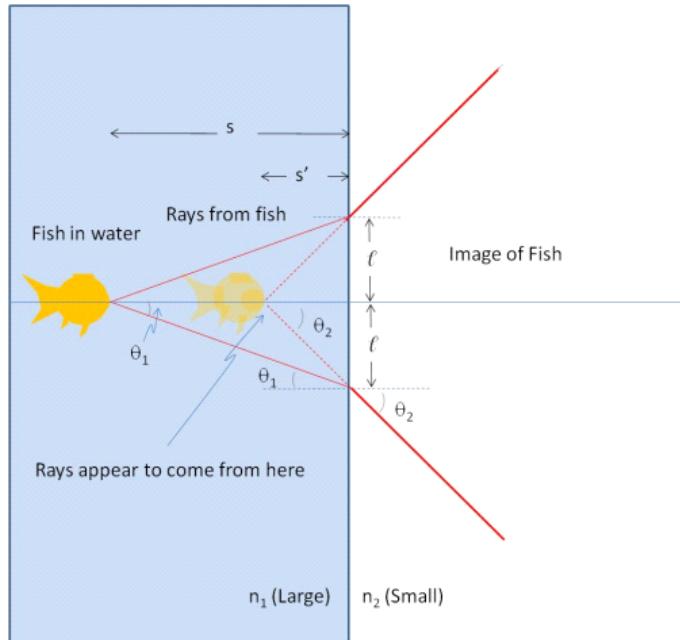
We can see that the picture is made from every point on the fish being “imaged” to a point on the retina. We collect the rays leaving every point on the fish, and bring them to corresponding points on the retina to make the picture.

It will take us a few lectures to see exactly how this is done by the lens system in the eye, but as a first step, let’s consider the fish tank, itself. Put the fish back in the tank and look at it.



Rays still come from the fish. But we now know that the change from a slow material to a fast material will bend the light. These bent rays are collected by our eyes, and the picture of the fish is formed on the retina just as before. But our eyes interpret the light as though it went in straight lines with no bends (dotted lines in the last figure). Our mind is designed to believe light travels in straight lines, so our mind tells us there is a fish, but that the fish head (and every other part of the fish) is closer than it really is. We call this apparent fish at the closer location an image of the fish, because this is where we think the diverging set of rays come from that form the fish pattern.

The next figure shows the details of the rays leaving a dot on the fish head



The dot on the fish head is our object for this set of rays. The distance from the fish-head dot and the edge of the water/air boundary is called the *object distance* and is given the symbol s .

The distance from the image of the fish-head dot to the edge of the water/air boundary is called the *image distance* and is given the symbol s' . Note that this is not a derivative, it is just a distance like s , because it appears to be where the rays come from, but it is a different distance because of the refraction of the rays. So to make it look different we put a prime mark on it.

Do this math, you will circle back to this

We can find where the image is (s') knowing s . We can see from the figure that

$$\begin{aligned}\ell &= s \tan \theta_1 \\ &= s' \tan \theta_2\end{aligned}$$

so

$$s \tan \theta_1 = s' \tan \theta_2$$

or

$$s \frac{\tan \theta_1}{\tan \theta_2} = s'$$

from Snell's law, we know that

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1}$$

Usually we can take the small angle approximation. This would limit our analysis to rays that are near the central axes. Let's call this central axis the *optics axis* and the rays that are not too far away from this axis *paraxial rays*. Then for our small angles we can write

$$\tan \theta_i \approx \sin \theta_i$$

so

$$s \frac{\tan \theta_1}{\tan \theta_2} \approx s \frac{\sin \theta_1}{\sin \theta_2} = s \frac{n_2}{n_1} = s'$$

and we have the image distance

$$s' = s \frac{n_2}{n_1}$$

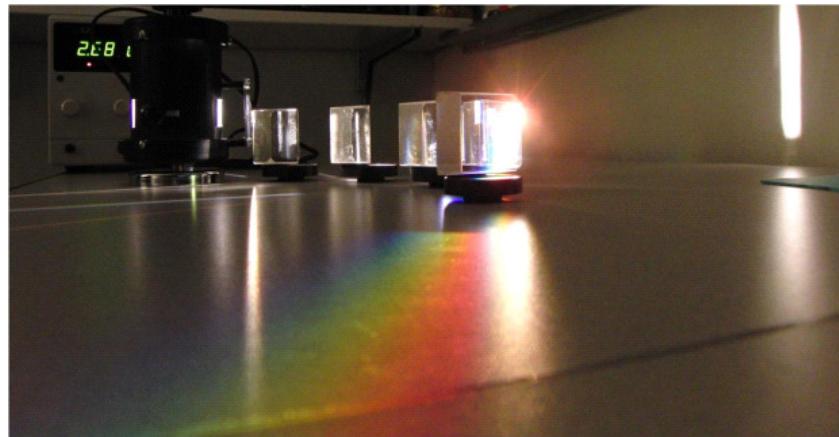
This is not so useful unless you have some burning need to know where your fish are in a tank. But we now have the vocabulary to discuss the larger problem of how a lens works, which we will take up next time.

15 Dispersion and Thin Lenses

Fundamental Concepts

- Index of refraction is wavelength dependent
- We can describe the operation of thin lenses using three easy-to-draw rays.

Dispersion



Question 223.15.1

Who hasn't played with a prism? We immediately recognize a rainbow. But why does the prism make a rainbow? The secret lies in the nature of the refractive index.

Notice that in the figure, the index of refraction depends on wavelength. This means that as light enters a material, different wavelengths will be refracted at different angles.

White light is not really a color of light. White light is made up of many colors—in fact, all the colors of the rainbow!⁹ Thus white light is pulled apart by refraction into a

⁹ Ah, but some light sources fool us. As long as there are the right amounts of red, green, and blue, we can

rainbow. This process is called *dispersion*. The reason is that for different wavelengths of light it is more likely for the light waves to be absorbed and re-emitted than for other wavelengths. This has to do with the spacing of the atoms relative to the wavelength, and it has to do with the electron structure of the material. Here is a graph that shows the index of refraction for some materials as a function of wavelength.

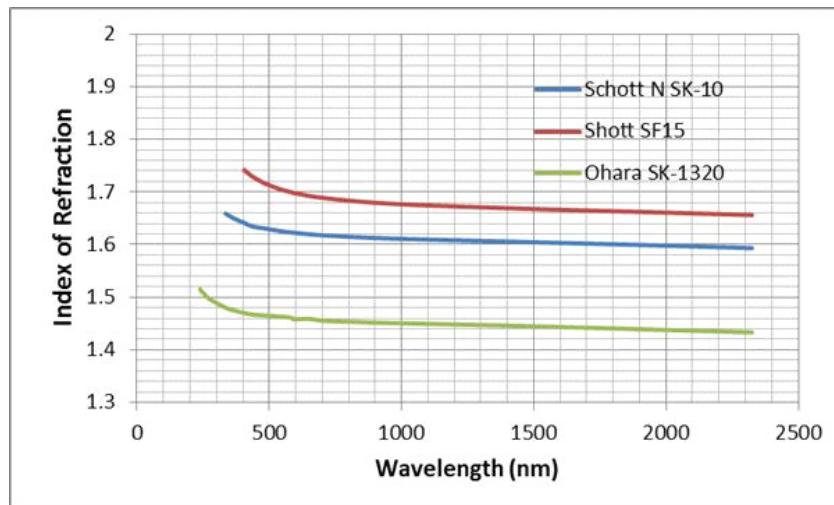
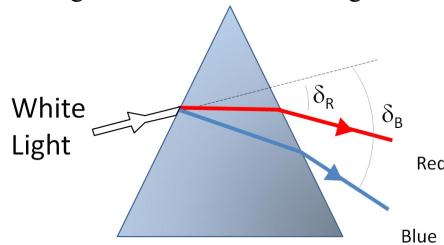


Figure 15.10. Index of refraction as a function of wavelength (Ohara optical glass <http://www.oharacorp.com/fused-silica-quartz.html> data and Schott optical glass data http://www.uqgoptics.com/materials_glasses_schott.aspx)

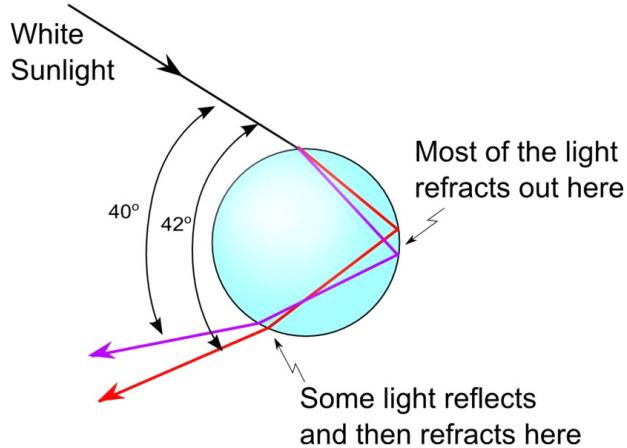
The graph tells us that blue light bends more than red light for these materials.



Question 223.15.2 We call the change in direction measured from the original direction of travel, δ , the *angle of deviation*. The colors we can see are called the visible spectrum.

think the light is white. Fluorescent lights and LED lights do this, and the lack of a full spectrum of light explains why plants don't grow well under fluorescent lights and some LED lights.

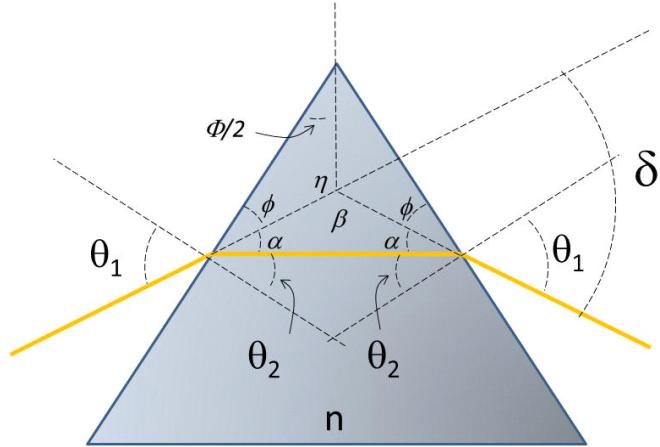
Let's look at a natural rainbow. The dispersion is caused by small droplets of water. The white sunlight enters the drop and is dispersed. It bounces off the back of the drop and then leaves the drop, again being dispersed. Red light leaves the drop at about 42° from its input direction, and blue light leaves at about 40° .



Calculation of n using a prism

Let's do a problem using the idea of dispersion. Let's find the index of refraction of a material. Suppose we make a prism as shown¹⁰. We know the apex angle of the triangle, Φ , and can measure the exit angle δ . In terms of these two variables, what is n ?

¹⁰ In this problem, we have carefully arranged the light so it goes horizontally across the prism. This is not always the case—and it is not usually the case in the problems in the homework.



Our strategy should be to use Snell's law

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

If we can find θ_1 and θ_2 in terms of Φ and δ then we can solve for the index of refraction of the material (we know $n_1 \approx 1$). Using the notation indicated in the figure, we choose θ_1 such that the interior ray is horizontal.¹¹

$$\theta_1 = \theta_2 + \alpha$$

$$\delta = 180 - \beta$$

$$180 = \beta + 2\alpha$$

Then

$$\delta = \beta + 2\alpha - \beta$$

or

$$\delta = 2\alpha$$

and

$$\alpha = \frac{\delta}{2}$$

¹¹ WARNING! in the homework problems you can't make the same assumptions!

$$90 = \alpha + \theta_2 + \phi$$

and

$$180 = \Phi + 2\alpha + 2\phi$$

or

$$90 = \frac{\Phi}{2} + \alpha + \phi$$

Then

$$\begin{aligned}\alpha + \theta_2 + \phi &= \frac{\Phi}{2} + \alpha + \phi \\ \theta_2 &= \frac{\Phi}{2}\end{aligned}$$

We can put these in our equation for θ_1

$$\begin{aligned}\theta_1 &= \theta_2 + \alpha \\ &= \frac{\Phi}{2} + \frac{\delta}{2} \\ &= \frac{\Phi + \delta}{2}\end{aligned}$$

We now know θ_1 and θ_2 . Now we can use Snell's Law to find n

$$\begin{aligned}\sin(\theta_1) &= n \sin(\theta_2) \\ \sin\left(\frac{\Phi + \delta}{2}\right) &= n \sin\left(\frac{\Phi}{2}\right)\end{aligned}$$

then

$$n = \frac{\sin\left(\frac{\Phi + \delta}{2}\right)}{\sin\left(\frac{\Phi}{2}\right)} \quad (15.1)$$

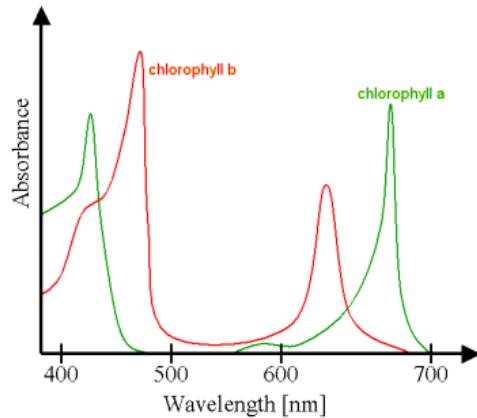
This gives a value for the index of refraction, but it would be better to repeat the analysis for several wavelengths. The resulting values for n can be combined into a n vs. λ curve like the one shown in figure 15.10.

Filters and other color phenomena

We have assumed without proof, that white light is made up of all the colors of the rainbow. Diffraction gratings were pretty good hints that this is true. Now that we know how prisms work, we have additional evidence of this. But knowing that white light is made a superposition of waves of different wavelengths, we should ask why a red shirt is red, or why passing light through a green film makes the light look green as it leaves.

Both of these phenomena are examples of removing wavelengths from white light.

In the case of the red shirt, the red dye in the cloth absorbs all of the visible colors except red. The red is reflected, so the shirt looks red. The filter is much the same. The green pigment in the film causes nearly all visible colors to be absorbed except green. So only green light is transmitted. This is why leaves are green. Chlorophyll absorbs red and blue wavelengths, so the green is reflected or transmitted.

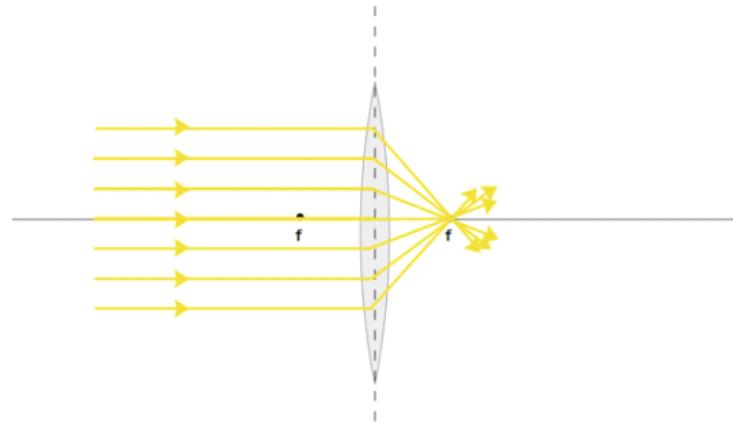


chlorophyll spectrum (Public Domain image courtesy Kurzon)

Knowing the nature of white light, we can start to understand lens systems and their challenges.

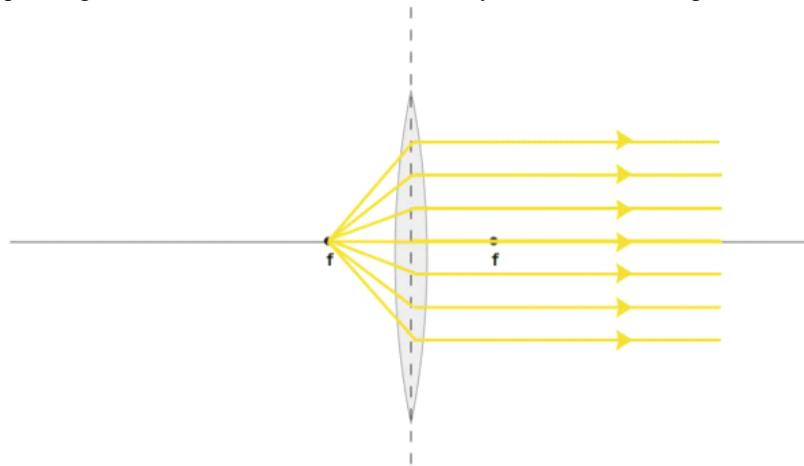
Ray Diagrams for Lenses

Before we do a lot of math to describe how lenses work, let's think about our early childhood experiences. You may have burned things with a magnifying glass. Using the idea of a ray diagram, here is what happens.

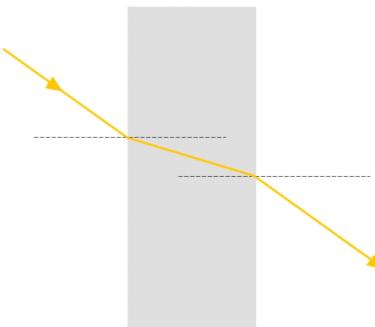


The rays from the Sun come from so far away that they are essentially parallel. We know that these rays come together to a fine point that can start a fire. The point where these rays converge is important to us. We will call this the *focal point*.

Knowing that the light will follow the same paths either direction, we would expect that if we put a light source at the focal distance, the rays should come out parallel.

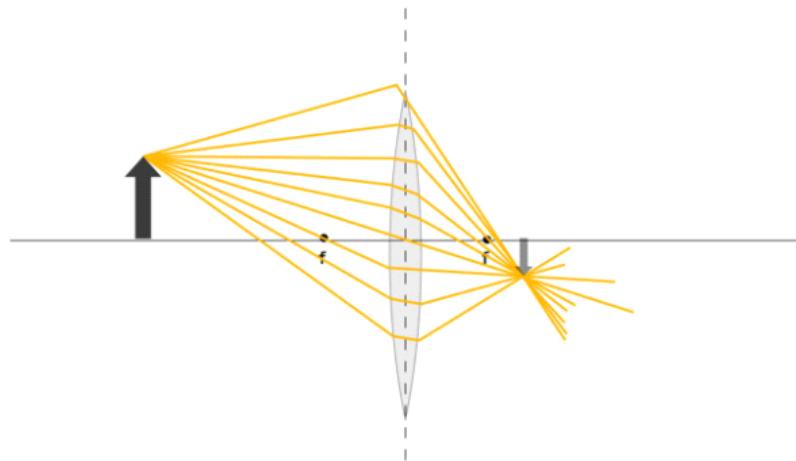


We need one other bit of information, to understand lenses. We have seen this case before.

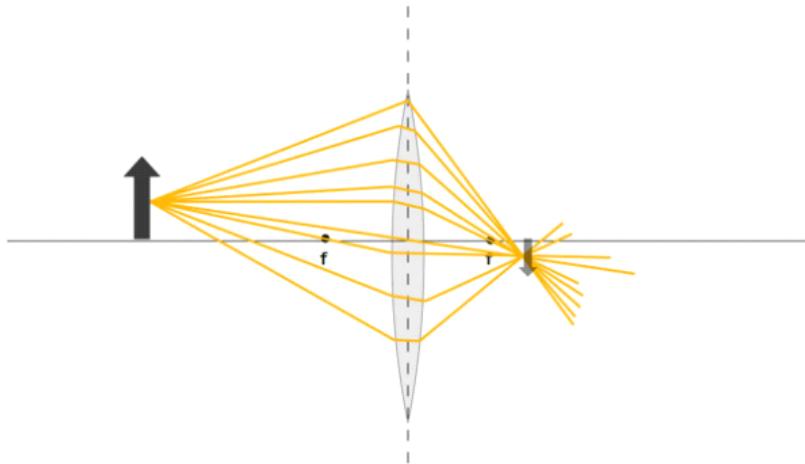


A flat block does refract the light, but when the light leaves the block it is only displaced, it retains the original direction. We will use these three situations to describe what happens when light travels through a lens.

We know that for every point on the object, we get millions of reflected rays that diverge. The lens must collect these rays together to form the corresponding point on the image.

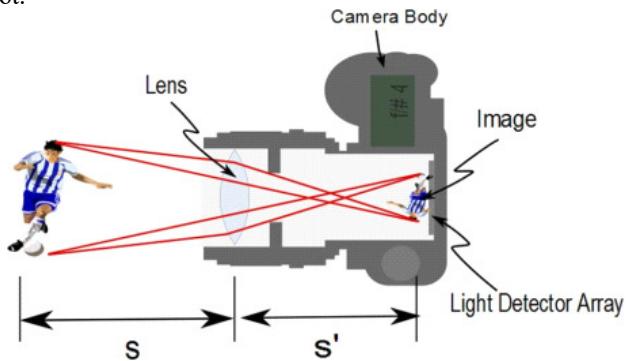


In the figure, the object is an upright arrow. We suppose that the arrow either glows, or that light is reflecting off the arrow. The arrow is a diffuse reflector, so the light bounces off in all directions. In the figure, you see light bouncing off the tip. Of course, this happens for every point on the image. Here is another drawing with light bouncing off the middle of the arrow.



But we usually pick the top of the object. If we place the bottom of the object on the optic axis, the bottom of the image will also be on the optic axis. So knowing where the bottom of the image is, and finding the top of the image gives a pretty good idea of where the rest of the image must be. So we will draw diagrams for the top of the object to find the top of the image.

But suppose this is not true? For example, when we use a camera, we do not align the optical system so the bottom of the subject is in the middle of the lens, on an axis, before we shoot.

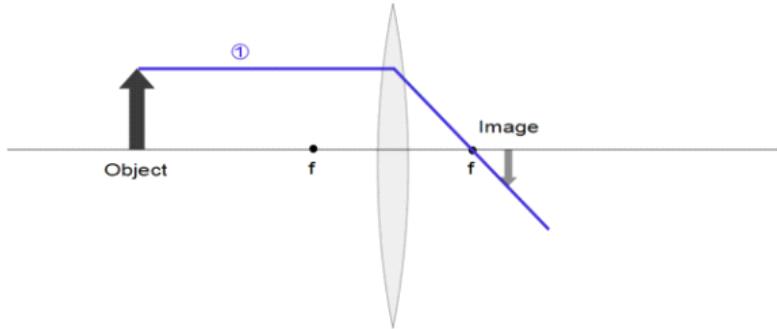


We can, of course, trace some rays for the bottom of the image as well as for the top in this case and find the location of the bottom of the image. The middle of the image will still be in between the top and the bottom.

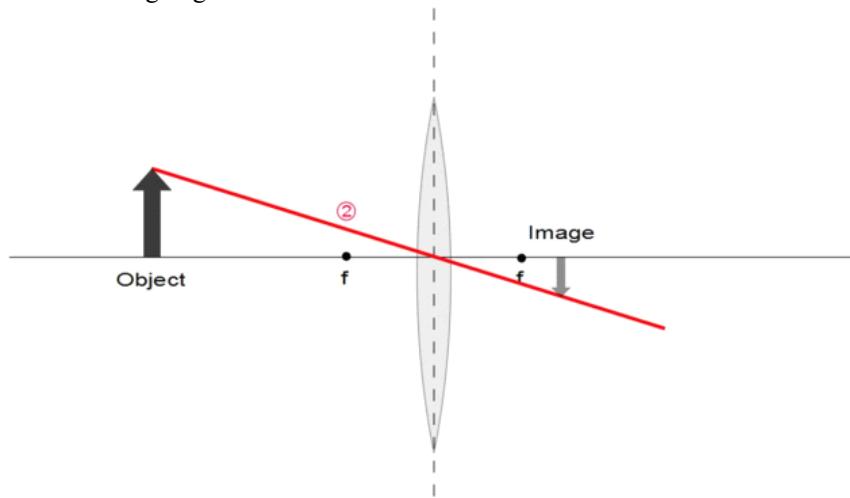
Notice that we said that light bounced off the arrow in all directions, but we did not draw all the rays going in all directions. Drawing millions of rays is impractical, and fortunately, not needed. We instinctively only drew rays that headed toward the lens. Any ray that does not head toward the lens won't take part in forming the image created

by the lens. But could we make due with even less rays?

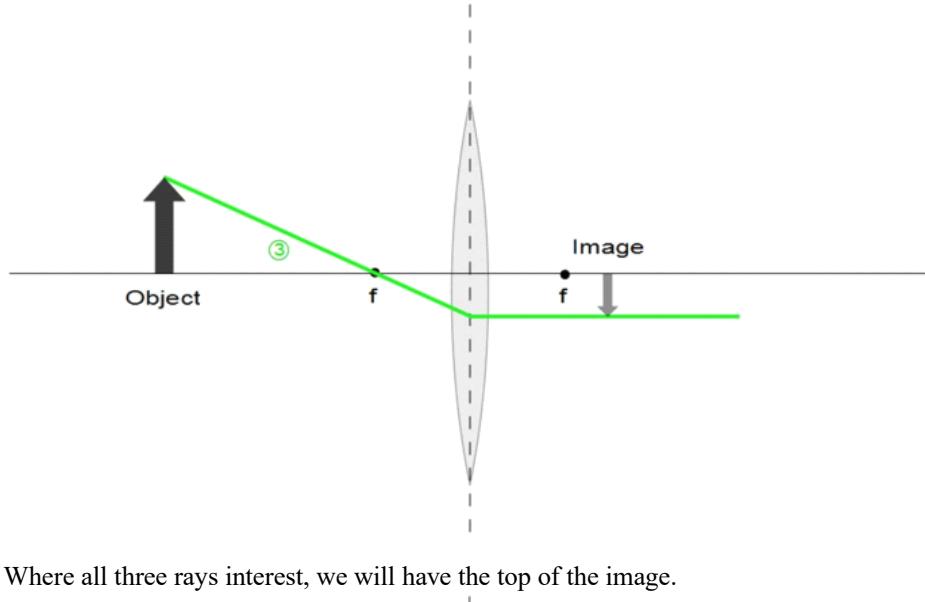
It turns out that we can choose three simple rays that leave the top of the object, and see where these rays converge to form the top of the image. Let's start with a ray that travels from the top of the object and travels parallel to the optic axis. We recognize this ray as being like one of the rays from the Sun. It comes in parallel, so it will leave the lens and travel through the focal point.



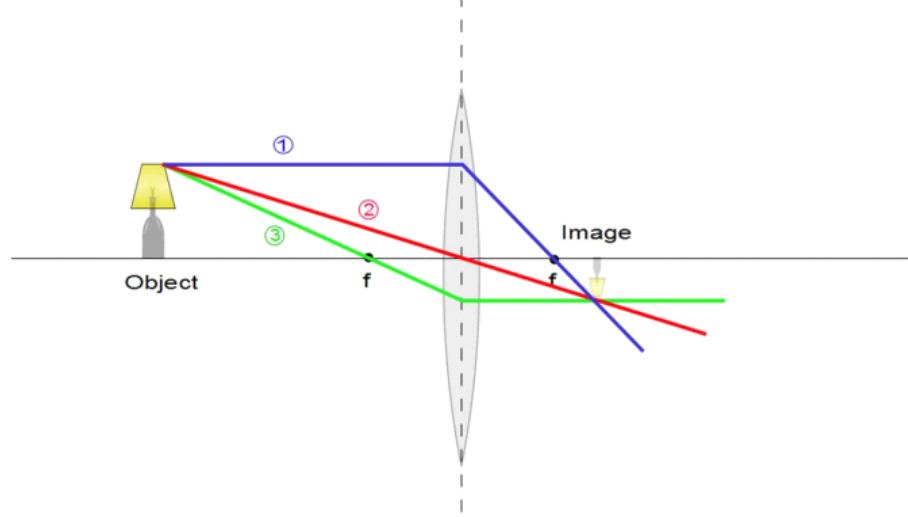
For a second easy ray, let's take the case that is like our flat block. Near the center of the lens, the sides are nearly flat. So we expect that the ray will leave in about the same direction as it was going before it struck the lens.



Two rays are really enough to determine where the top of the image will be, but there is a third ray that is easy to draw, so let's draw it to give us more confidence in our answer. That ray is one that leaves the top of the object and passes through the focal point on the object side of the lens. This situation we also recognize. This ray will leave the lens parallel to the optic axis.



Where all three rays interest, we will have the top of the image.



Because the rays come together or *converge* we call a lens like this a *converging lens*. Notice that in this case, the image is upside down. That is normal. Also notice that it is smaller than the object. We say that the image is magnified, which may seem a little bit strange. But in optics, a magnification of greater than one means that the image is bigger than the object. This is like a movie projector that makes a large image of a small film segment. The magnification can be equal to one, meaning the object and image are the same size. And finally, the magnification can be less than one. This means that the image is smaller than the object. This is a convenient definition, because then we can

use the same equation to describe all three situations.

$$m \equiv \frac{\text{Image height}}{\text{Object height}} = \frac{h'}{h}$$

where h is the object height, and h' is the image height. It turns out that we can also write the magnification of a lens in terms of the object distance, s and image distance s' (see more on this below).

$$m = -\frac{s'}{s}$$

Notice the negative sign. By convention (meaning physicists got together and voted on this) we say that an upside down image has a negative magnification. You just have to memorize this, there is no obvious reason for this except it is mathematically convenient.

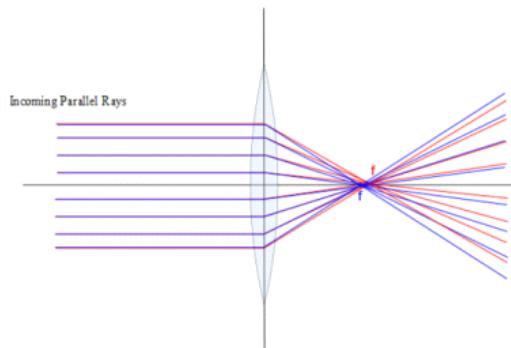
Question 223.15.4

Thin Lenses

It is time to introduce another approximation Suppose the lens is very thin. Then ray number 2 would travel through the lens with no deviation at all. This is sometimes a good approximation, and will make the math easier, so for this class we will often use it. But there are times when it really does not work, so in practice you have to be careful. PH375 goes beyond the thin lens formulation.

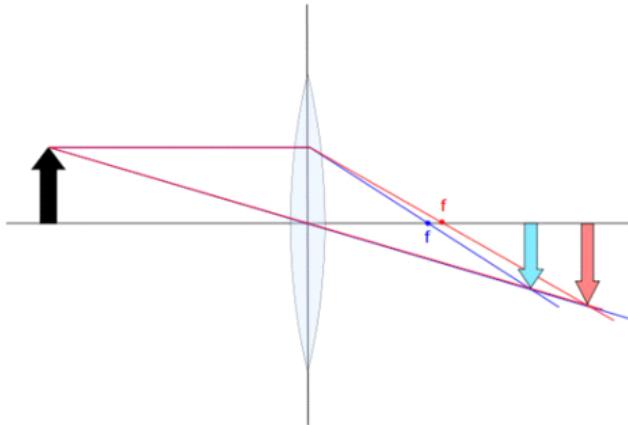
Question 223.15.3

We should pause to realize that our new understanding of Snell's law tells us that we have a problem with our lenses. The index of refraction is wavelength dependent. This means that different wavelengths will focus in different positions. Here is our light from the Sun again, but note that I drew blue light and red light only.



Having removed all the other colors, we can see that the blue light focuses nearer to the lens than the red light. This is because the index of refraction for the blue light is larger. Each visible wavelength will focus somewhere in between these two (except for purple,

of course). When we make an image, this means that we get multiple images of our object, one in each color. Usually the images overlap, so we end up with a colored blur.



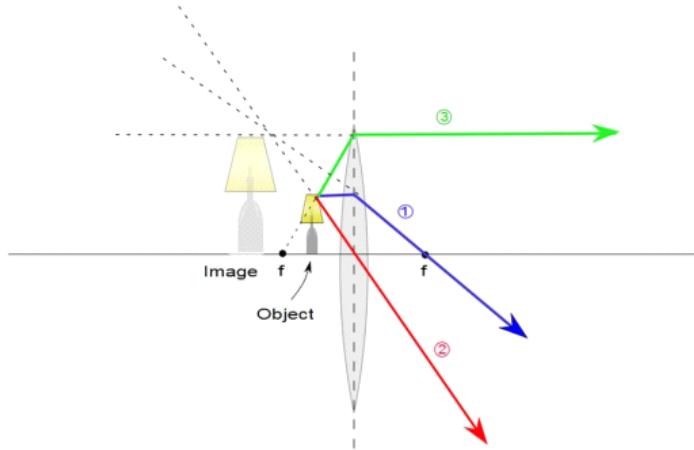
This problem is called Chromatic Aberration. We can fix this by using a combination of lenses.



where each lens has a different index of refraction. The converging lens is designed to form the image, while the diverging lens (a term we explain below) realigns all the colors.

Virtual images

Lets take another case and draw a ray diagram. This time let's place the object closer than the focal distance. This is the case when we use a lens as a magnifying glass. The rays will look like this.



Notice that these rays never converge! We won't get an image that could project on a paper. But we know that there is an image, we can look through the lens and see it! And that is the key. The image does not really exist. We look through the lens, and our mind interprets the diverging rays coming from the lens as though they had only traveled in straight lines. If we extend these rays backwards along straight lines, they appear to come from a common point. This is the point they would have had to have come from if there were no lens. Because our brain believes light travels in straight lines, we believe we see an image at this location. But no light really goes there! Because this image is not really made from light diverging from this position, we recognize this as a *virtual image*. The image we formed before that could be projected on a screen is called a *real image*.

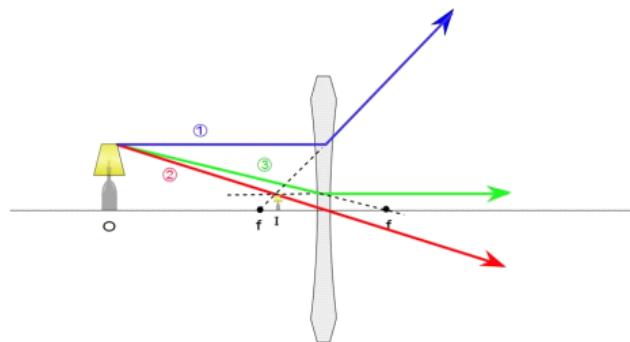
By convention, we say the distance, s' , from the lens to the virtual image has a negative value.

Diverging Lenses

So far our lenses have only been the sort that work as magnifying glasses. We call these *converging lenses*. These lenses are fatter in the middle and thinner on the edges. Because of this they are sometimes called *convex lenses*. By convention, we say the focal distance for this type of lens is positive. For this reason, they are often called *positive lenses*.

But what if we make a lens that is thinner in the middle and thicker on the edges. We can call this sort of lens a *concave lens*, and we will give it a negative focal length by convention, so we can also call it a *negative lens*. But what would this lens do? If we

think about our three rays and Snell's law, ray 1 won't be bent toward the optic axis for this type of lens. In fact, if we observe an object through this lens, ray number 1 will appear to come from the focal point. Ray number 2 will still go through the middle of the lens, and if the lens is thin enough, ray 2 will pass through undeviated.



finally ray three will go as if it were aiming for the far focal point, but it will hit the lens and leave parallel to the optic axis. From the figure we see that these three rays will never converge. We expect they will form a virtual image. If we extend the rays backward as shown, we see that the extensions all meet at a point. The rays leaving the lens appear to come from this point. This is the location of the virtual image.

You might wonder what good such a lens could do, but we will find that this type of lens is used to correct vision for nearsighted people.

16 Image Formation

Last lecture we learned how to find an image location graphically, now let's do it algebraically.

Fundamental Concepts

- A curved interface between two media can cause light rays to cross
- The lens-maker's formula is given by $\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$
- The thin lens formula is given by $\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$
- The sign of quantities that go into the lens-maker's equation and the thin lens formula are determined by a sign convention.

Thin lenses and image equation

Question 223.165.1

In this lecture, we will work toward understanding the equations that allow us to solve for the image location given the object location for a thin lens. Let's start by thinking of a special case for refraction. A circular or spherically curved surface on a very large piece of glass. We will assume that the piece of glass is semi-infinite, but all it has to be is very large.

We can call this a semi-infinite bump of glass.

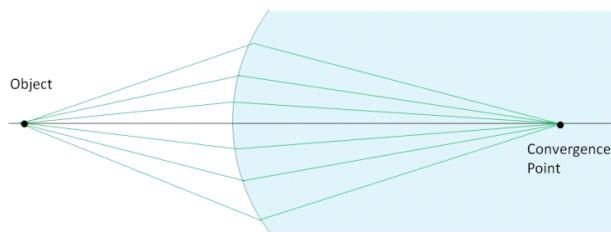


Figure 16.11.

Take a point object that either glows, or has rays of light reflecting from it. The rays leave the object and reach the surface of the glass. The rays will refract at the surface. Each bends toward the normal, but because of the curvature of the glass, the rays all converge toward the center. We can identify this convergence point as the image of the point object. Since our object is a point, so is our image.¹² Of course we could make an extended object out of many points, and then we would have many image points as well.

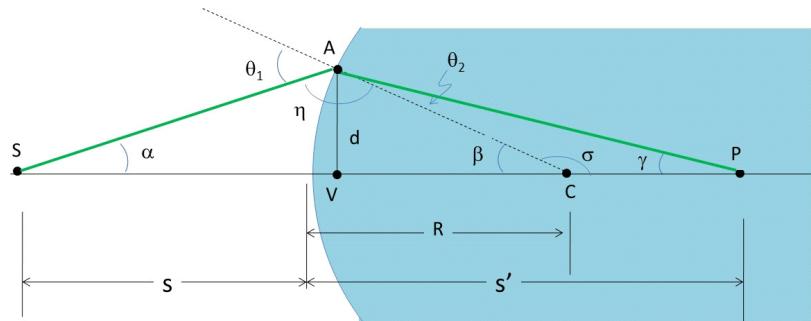
At the surface we can find the refracted angles using Snell's law

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

We will again use the small angle approximation. Then θ_1 and θ_2 are small and none of the rays are very far away from the axis. This is our paraxial approximation. Snell's law becomes

$$n_1 \theta_1 = n_2 \theta_2$$

Let's try to see where the image will be using Snell's law.



Using the more detailed figure above, we observe triangles SAC and PAC . We recall that for triangle SAC the top angle labeled η , plus θ_1 must be 180.

$$180^\circ = \theta_1 + \eta$$

or

$$\eta = 180^\circ - \theta_1$$

We also know that the sum of interior angles must equal 180. So for triangle SAC we know

$$180^\circ = \eta + \alpha + \beta$$

$$180^\circ = 180^\circ - \theta_1 + \alpha + \beta$$

then

$$\theta_1 = \alpha + \beta$$

¹² Within the limits of diffraction. So it is really a small circle of light. But it's mostly a point.

Likewise, from triangle PAC ,

$$180^\circ = \sigma + \theta_2 + \gamma$$

and

$$180^\circ = \beta + \sigma$$

so

$$\sigma = 180^\circ - \beta$$

and

$$180^\circ = (180^\circ - \beta) + \theta_2 + \gamma$$

which reduced to

$$\beta = \theta_2 + \gamma$$

then,

$$\theta_2 = \beta - \gamma$$

and we can write our paraxial Snell's law as

$$n_1\theta_1 = n_2\theta_2$$

$$n_1(\alpha + \beta) = n_2(\beta - \gamma)$$

$$n_1\alpha + n_1\beta = n_2\beta - n_2\gamma$$

$$n_1\alpha + n_2\gamma = n_2\beta - n_1\beta$$

$$n_1\alpha + n_2\gamma = \beta(n_2 - n_1)$$

Looking at the figure. We see that d is a leg of three different right triangles (SAV , ACV , and PAV). The ray in the figure is clearly not a paraxial ray. If we use an actual paraxial ray, then the point V will approach the air-glass boundary. When this happens, then $SV = s$, $VC = R$, and $VP = s'$. So we can write

$$\begin{aligned}\tan \alpha &\approx \alpha \approx \frac{d}{s} \\ \tan \beta &\approx \beta \approx \frac{d}{R} \\ \tan \gamma &\approx \gamma \approx \frac{d}{s'}\end{aligned}$$

so our Snell's law becomes

$$\begin{aligned}n_1\alpha + n_2\gamma &= \beta(n_2 - n_1) \\ n_1\frac{d}{s} + n_2\frac{d}{s'} &= \frac{d}{R}(n_2 - n_1)\end{aligned}$$

We can divide out the common factor, d .

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{(n_2 - n_1)}{R} \quad (16.1)$$

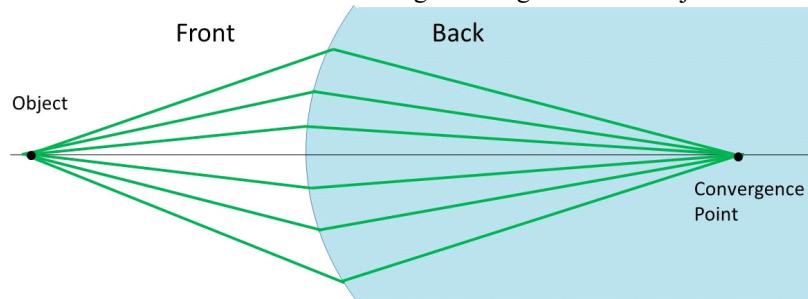
We can use this formula to convince ourselves that no matter what the angle is (providing it is small), the rays will form an image at P . So all the rays in figure (16.11) will converge at P .

Real images will be inside the glass for our example. This may seem a problem, but we will fix this with non-infinite lenses soon. And for the case of our eyes, this is exactly what happens. We have fluid (sort of a jelly) in our eyes, and the image is formed in the fluid. The curved surface is our cornea (the spot where your contacts go).

Physicists got together and decided on a mathematical system of signs to make the math easier and consistent. We have called such a scheme a sign convention already. We started collecting parts of this system last lecture. Let's write it all out in a table so we can use it in today's lecture. Here is the convention for the case of a curved semi-infinite surface.

Quantity	Positive if	Negative if
Object location (s)	Object is in front of surface	Object is in back of surface (virtual object)
Image location (s')	Image is in back of surface (real image)	Image is in front of surface (virtual image)
Image height (h')	Image is upright	Image is inverted
Radius (R)	Center of curvature is in back of surface	Center of curvature is in front of surface

where the “front” of the lens is the side that gets the light from the object.



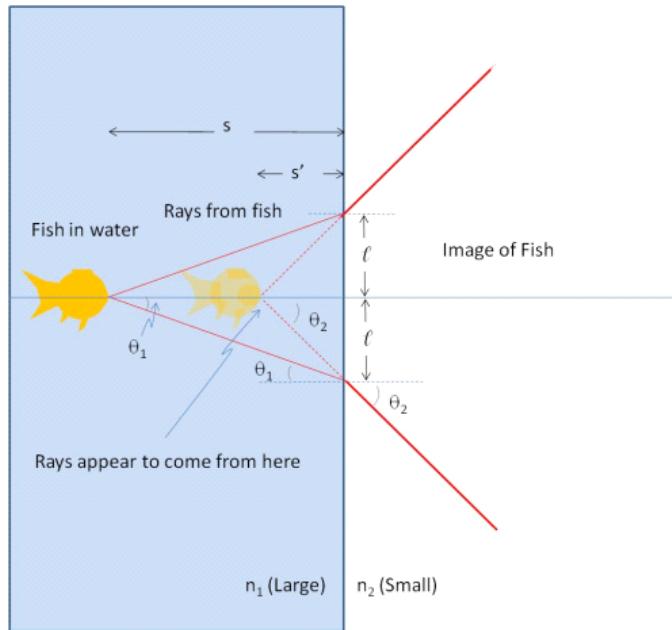
Question 223.165.2

Question 223.165.3

Question 223.165.4

We could go through the entire derivation and switch the indices of refraction (in effect, go along the same path, but going backwards). It turns out we get the same equation. The light is bending the other way as it travels the path, but the equation will be the same. So our equation describes light entering a piece of glass, or light leaving a piece of glass.

Flat Refracting surfaces



Let's return to our fist tank. The fish tank has an interface, but it is flat. Can we use our equation (16.1) to describe this?

The answer is yes, if we let $R = \infty$. This makes sense for a flat surface. If we have an infinitely large sphere, then our small part of that spherical surface that makes up the fish tank wall will be very flat.

Then

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{(n_2 - n_1)}{\infty}$$

or

$$\frac{n_1}{s} + \frac{n_2}{s'} = 0$$

we see that

$$s' = -s \frac{n_2}{n_1}$$

This is what we got before for this case, except before we just got the distance, and now we have included the effects of our sign convention. The negative sign means that the image is in front of the surface. By "in front" we always mean to follow the light from

the source (fish) to the optical boundary. This boundary is the water/air boundary of the tank, so the fact that our image is in the water means that our image is in front of the optical boundary. As we know, this means the image is virtual.

Thin Lenses

Question 223.165.5

Lets' find an equation for a lens made from sections of spherical surfaces once more. But this time, let's let it be more practical and not make the "lens" semi-infinite. We will need to deal with two sides of the lens because (usually) both will be curved.

We found that for refraction

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{(n_2 - n_1)}{R}$$

but we did this for a spherical bump on a semi-infinite piece of glass. For this problem let's make a few assumptions:

- We have two spherical surfaces, with R_1 and R_2 as the radii of curvature
- We have only paraxial rays
- The image formed by one refractive surface serves as the object for the second surface
- The lens is not very thick (the thickness is much smaller than both R_1 and R_2)

The answer we will get is quite simple

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad (16.2)$$

where

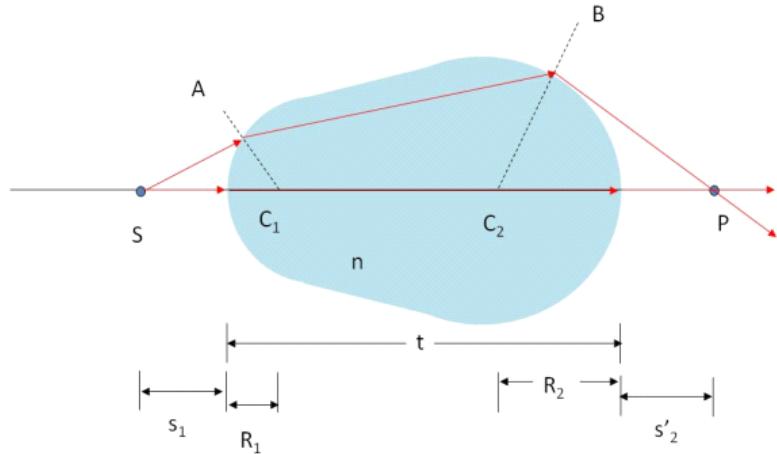
$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (16.3)$$

but to appreciate what it means, lets find out where it comes from.

Derivation of the lens equation

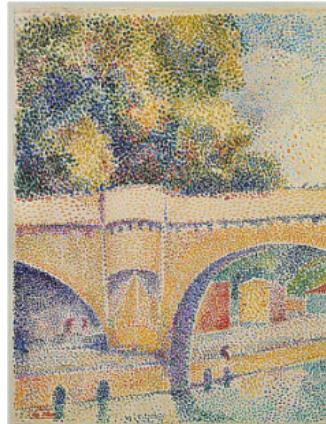
Let's find the thin lens formula. Really, we could just assume the formula and be fine, but we are going through the derivation because it will help teach us how to deal with multiple surfaces in an optical system. Telescopes and microscopes all have multiple surfaces. So this will help us understand how they work.

Consider the optical element in the figure below. Notice that our object is a dot, so our image will also be a dot.



Question 223.165.6

By now you have realized that this is not as boring as it sounds if we consider any object can be considered as a collection of dots.



The Pont Neuf by Hippolyte Petitjean. Petitjean was a Pointillist, one who painted with dots of paint instead of continuous application of paint. This illustrates the thought that all objects can be thought of collections of small points that reflect or emit light. So we can consider an optical system by considering individual points of light and how the system reacts to those points of light. (Image in the Public Domain)

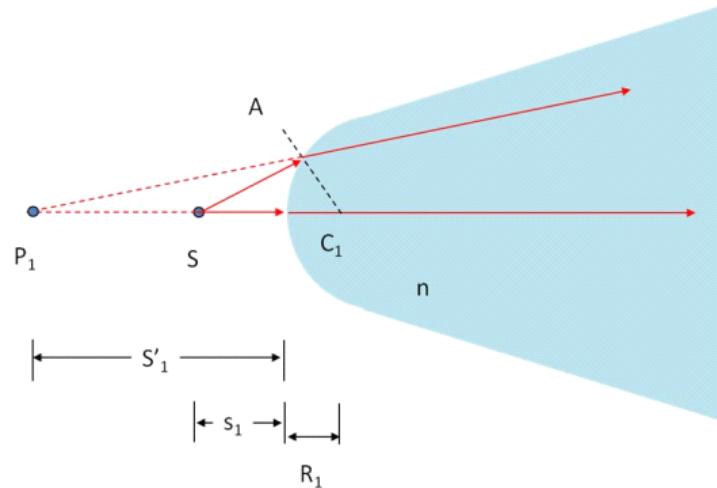
So we can consider anything as a collection of dots, and work out our formulas for a dot (because it is easier to think about just one dot at a time).

Light enters at a spherical surface on the left hand side. We use a point object located at S on the principal axis, and trace two rays. The ray along the principal axis crosses each spherical surface at right angles, and therefore travels straight through the optic. The second ray hits the first spherical surface at point A . It is refracted and travels to

point B . It is again refracted and travels toward the principal axis, crossing at P . The image location is the intersection of these rays, so we have an image at P .

Lets study the surfaces separately

Surface 1:



We treat surface 1 as though surface 2 did not exist. After all, the light does not know about surface 2 as it hits surface 1.

By considering surface 1 on its own, we have just our semi-infinite bump problem, so we know that

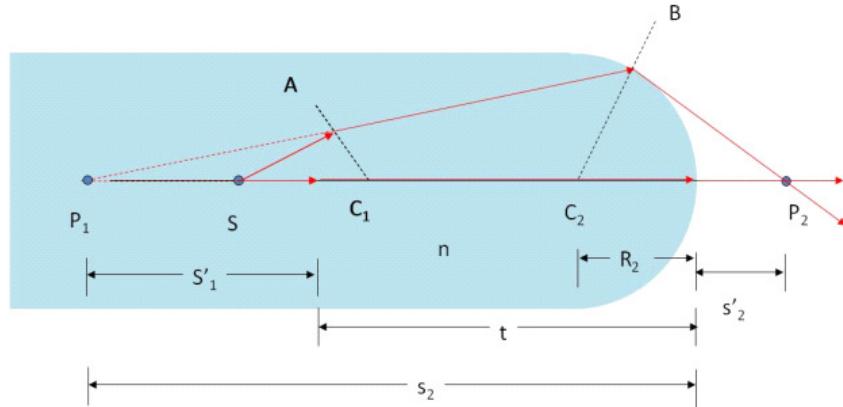
$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{(n_2 - n_1)}{R}$$

We can consider $n_1 = 1$ and $n_2 = n$ for an air-glass interface and noting that s'_1 is negative by our convention. Then

$$\frac{1}{s_1} - \frac{n}{s'_1} = \frac{(n-1)}{R_1} \quad (16.4)$$

Note that our rays are *not* converging in the glass. We can find the image formed by this surface 1 of our lens by tracing the diverging rays backward as we did for the fish tank or magnifying glass. The image formed from the first side of the lens is virtual.

Surface 2: Now consider the second surface.



The second surface sees light diverging as though it came from a semi-infinite piece of glass with the origin at P_1 . The virtual image formed by surface 1 serves as the object for surface 2 because the diverging light is just the same pattern as if there were a light source at P_1 . The distance from P_1 to surface 2 is

$$s_2 = s'_1 + t$$

We again use our refractive equation for a semi-infinite bump

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{(n_2 - n_1)}{R}$$

but we identify $n_1 = n$ and $n_2 = 1$. We have for surface 2

$$\frac{n}{s_2} + \frac{1}{s'_2} = \frac{(1 - n)}{R_2} \quad (16.5)$$

or

$$\frac{n}{s'_1 + t} + \frac{1}{s'_2} = \frac{(1 - n)}{R_2} \quad (16.6)$$

Now we take our thin lens approximation. Let $t \rightarrow 0$. Then equations 16.4 and 16.6 become

$$\frac{1}{s_1} - \frac{n}{s'_1} = \frac{(n - 1)}{R_1}$$

$$\frac{n}{s_1} + \frac{1}{s'_2} = \frac{(1 - n)}{R_2}$$

I would like a single equation that gives s'_2 in terms of s_1 . That is the form of the thin lens equation that we are looking for. Adding these two equations can give me such an equation.

$$\frac{1}{s_1} - \frac{n}{s'_1} + \frac{n}{s'_1} + \frac{1}{s'_2} = \frac{(n - 1)}{R_1} + \frac{(1 - n)}{R_2}$$

or

$$\frac{1}{s_1} + \frac{1}{s'_2} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

This equation is very useful. If we again let $s_1 = \infty$ (put the object at ∞ so the rays enter surface 1 parallel) we find

$$\frac{1}{s'_2} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

The spot where the rays gather if the object is infinitely far away is the focal point, f , so we can identify $s'_2 = f$ as the focal length of the optic in this special case.

Then we can identify

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

which is known as the *lens makers' equation*. It gives us a way to make a lens that will have a particular focal distance. You grind one side of the lens to have a radius of curvature R_1 and the other side to have radius of curvature R_2 . Then with index n , you will have the focal length you desire.

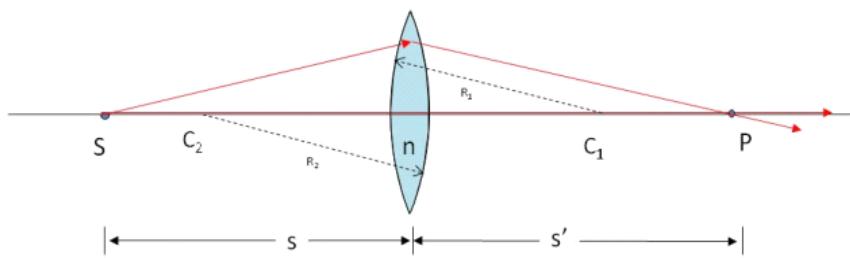
We have a relationship between the object distance in front of the lens, and the final image in back of the lens:

$$\begin{aligned} \frac{1}{s_1} + \frac{1}{s'_2} &= (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \\ &= \frac{1}{f} \end{aligned}$$

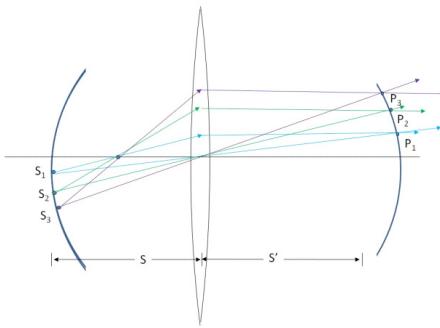
If we drop the subscripts (which we can do now that we let $t = 0$ since the internal distances for the inside points are not important) then.

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

This is called the *thin lens equation*. The resulting approximate geometry is shown below.



Of course any real object is made of lots of points, but each point is imaged in a corresponding point on the image



so, as we claimed earlier, our simple analysis explains the formation of actual images

Question 123.16.7

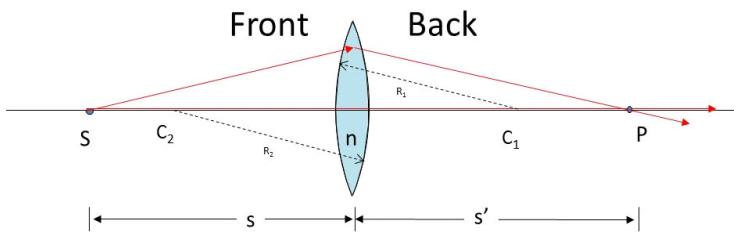
and not just point images.

Sign Convention

We need to add to our sign convention table a second radius, and the focal length.

Quantity	Positive if	Negative if
Object location (s)	Object is in front of surface	Object is in back of surface (virtual object)
Image location (s')	Image is in back of surface (real image)	Image is in front of surface (virtual image)
Image height (h')	Image is upright	Image is inverted
Radius (R_1 and R_2)	Center of curvature is in back of surface	Center of curvature is in front of surface
Focal length (f)	Converging lens	Diverging lens

Again the front surface is the surface that gets the light from the object.



Note that each radius has a sign. If the two radii are the same magnitude, it looks like

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

should be undefined (the focal length should be infinite) but usually that is not true

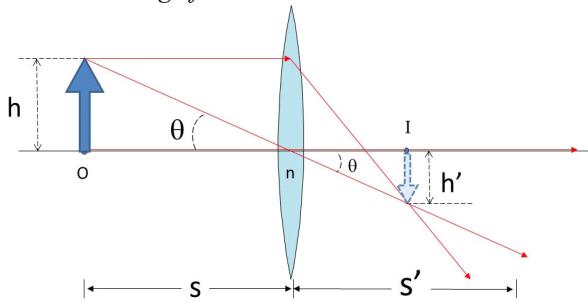
because either R_1 or R_2 will be negative.

Magnification

The image is not likely to be the same size as the object. We would like to have a quantity that tells us how big the image is. The measure of choice is the ratio of the two heights.

$$m = \frac{h'}{h} \quad (16.7)$$

where h is the object height and h' is the image height. Note that with our sign convention, if $m > 0$ then the image is upright, and if $m < 0$ the image is inverted (upside down). We call this ratio the *magnification* of the lens.



We can find an expression for the magnification in terms of s and s' . By observing the figure, and using the ray that goes right through the middle of the lens, we can see that

$$\tan \theta = \frac{h}{s}$$

and

$$\tan \theta = \frac{h'}{s'}$$

thus

$$\frac{h}{s} = \frac{h'}{s'}$$

then

$$\frac{s'}{s} = \frac{h'}{h}$$

which we can use to form a new equation for the magnification.

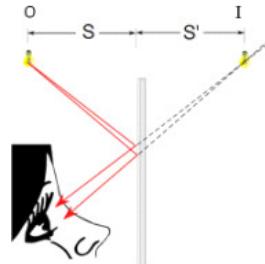
$$m = -\frac{s'}{s} \quad (16.8)$$

Images formed by Mirrors

Question 123.16.8

Question 123.16.9

All of us have looked in a mirror at some time. We know what to expect. We see an image of ourselves. To study mirrors we need to establish a sign convention and some standard notation

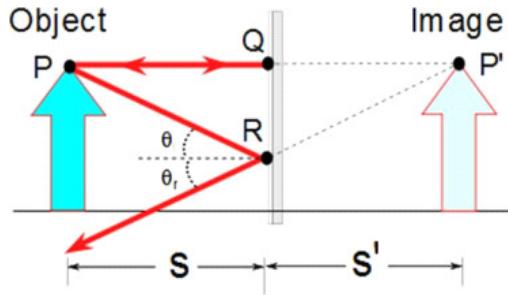


In the figure above, we have a person observing an object O in a mirror. The object is located at a distance s from the mirror. Just like with lenses, we will call this the *object distance*. The image appears to be located at a point I beyond the mirror a distance s' . This is the *image distance*.

It might be good to review how images are formed. Images are located at a point from which rays of light diverge *or at a point from which rays of light appear to diverge*. This only makes sense. If you remember how we see things, our eyes intercept rays of light diverging from an object. So if we can create a situation that makes rays diverge in the same way the object did, we will have an image of the object.

Mirrors create what we have called *virtual images* because the image appears to be created from diverging rays from behind the mirror, but if we look behind the mirror no rays exist at the image location (if they did exist, they would not make it through the mirror!).

Image from a Flat Mirror



Let's look at a simple image as shown in the figure above. The object (of course) is an arrow. We could trace all the rays that diverge from this object and build a very nice representation of the arrow¹³ but that would take time and computation power. We only really need to use two rays, and remember what the object looked like.

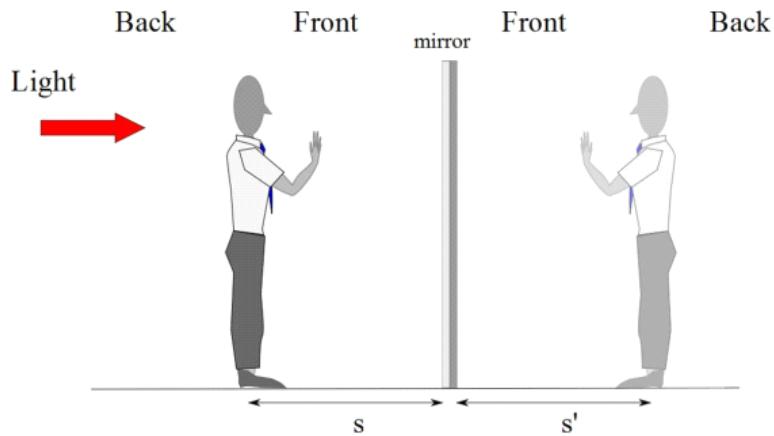
We pick one ray from the top of the arrow that travels straight to the mirror. This ray will travel a distance s and bounce back. We pick a second ray from point P that travels the path PR . This ray bounces off the mirror at an angle θ . So it appears that the tip of the arrow is at position P' and the rays from the tip appear to travel the paths $P'P$ and $P'R$. Again, our visual processing center in our brain interprets the rays as traveling in straight lines.

Mirror reversal

Question 223.16.10

Look into a mirror. Raise your left hand. Your image raises what appears to be a right hand. It looks like a mirror switches the left and right sides of the image. But lie sideways on the ground in front of the mirror and raise a hand. Your hand does not get inverted (and neither do your feet and head). What is happening? A flat mirror performs a front-back reversal. What this means is that, following the light direction, the object is positioned so that the back is encountered first, then the front, but in the image the front of the image is encountered first, then the back.

¹³ Ray tracing-based computer graphics actually does this—the way movies like *Toy Story* are made.

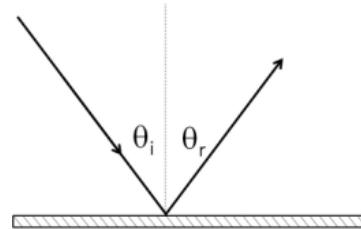


this has the effect of making it look like the left hand is raised when the object's right hand is raised.

Concave Mirrors

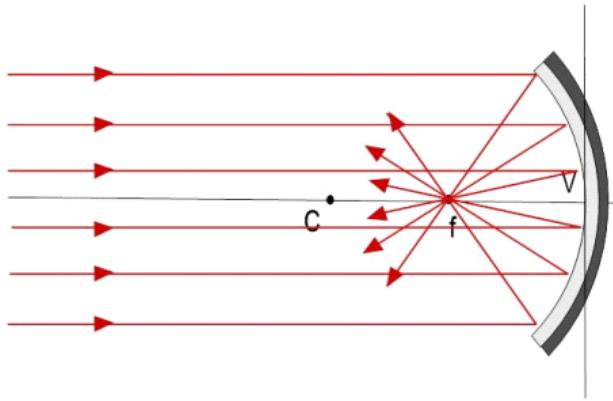
Question 223.16.11

Concave mirrors can form images. I'm sure you know that many telescopes are made with mirrors. We should see how this works. Let's proceed like we did for lenses, but looking at what happens when light strikes the surface of the new material, this time the mirror surface. First, let's look at rays that come from very distant objects so they enter parallel to the optic axis. We recall the law of reflection



$$\theta_i = \theta_r$$

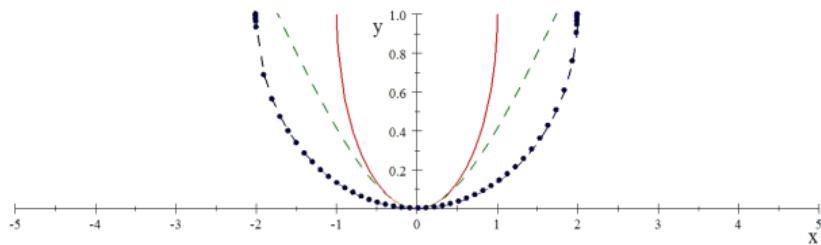
Armed with this, we can see what would happen. Each ray has a different normal due to the curvature of the mirror. The result is that the rays all meet at a spot on the axis.



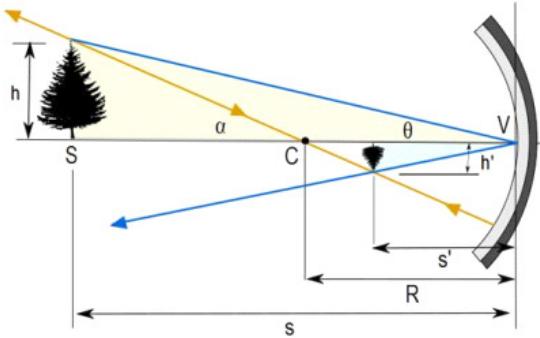
This is a focal point!

Paraxial Approximation for Mirrors

The correct shape of a mirror is more like a parabola, but parabolas are hard to machine or build. Spherical shapes are relatively easy. So we often see spherical mirrors just like we often see spherical lenses. This will work so long as we allow only rays that make small angles with respect to the principal axis. We can see why this works if we plot a sphere and a parabola (and a hyperbola). For small deviations from the center, the shape of the functions all look alike.



We would expect the reflections to be similar under these circumstances, so, if we meet the criteria for the paraxial approximation, our spherical mirrors should work. Note that when you need the entire mirror, say, in a communications antenna, you must do better than a spherical approximation to the correct shape for your mirror.



Like with our flat mirror, we will measure distances from the mirror surface (from point V). We can find the image location by again taking two rays. One convenient ray is the ray that passes through the center of curvature, C . This ray will strike the mirror surface at right angles and bounce back along the same path. Another convenient ray is the ray from the tip of the object to point V . This ray will bounce back with angle θ . Where these two reflected rays cross, we will find the image of the tip of our object (a tree this time, I got tired of imaging arrows). Knowing the shape of the object and that the bottom is on the axis, we can fill in the rest of the image.

We can calculate the magnification for this case. We use the gold triangle to determine that

$$\tan \theta = \frac{h}{s}$$

and the blue triangle to determine that

$$\tan \theta = \frac{h'}{s'}$$

so we have

$$m = \frac{h'}{h} = \frac{-s' \tan \theta}{s \tan \theta} = -\frac{s'}{s}$$

We want to indicate that the image is inverted by making it's sign negative. We recall that h' is negative if the image is inverted. So we added a negative sign to make this fit with our sign convention.

Mirror Equation

We can further exploit this geometry to get a relationship between s , s' , and R . Notice that

$$\tan \alpha = \frac{h}{s - R}$$

and that

$$\tan \alpha = \frac{-h'}{R - s'}$$

Then

$$\frac{h}{s - R} = \frac{-h'}{R - s'}$$

or

$$\frac{R - s'}{s - R} = -\frac{h'}{h}$$

We can use our magnification definition to replace h'/h

$$\frac{R - s'}{s - R} = \frac{s'}{s}$$

we perform some algebra

$$\begin{aligned} (R - s')s &= s'(s - R) \\ -s's + Rs &= ss' - Rs' \\ Rs + Rs' &= ss' + s's \\ Rs + Rs' &= 2ss' \\ \frac{Rs'}{Rs s'} + \frac{Rs}{Rs s'} &= \frac{2ss'}{Rs s'} \\ \frac{1}{s} + \frac{1}{s'} &= \frac{2}{R} \end{aligned}$$

This is called the *mirror equation*.

Focal Point for Mirrors

Now that we know the mirror equation, let's let s be very large. Then

$$\frac{1}{s'} \approx \frac{2}{R}$$

or

$$s' \approx \frac{R}{2}$$

Using the same logic as with the lens, we can identify this as the *focal point*, F and the distance s' in this case will be called the *focal length*, f . We see that

$$f = \frac{R}{2} \tag{16.9}$$

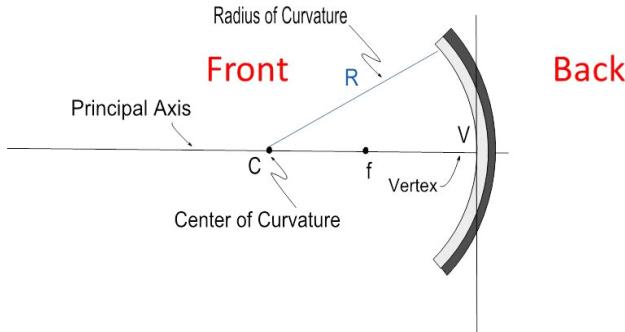
so we can write the mirror equation as

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{16.10}$$

For a mirror, the value of f does not depend on the mirror material (this is not true for refractive optics). Of course we have a sign convention, but it is similar to the convention for lenses. Here is the convention for mirrors.

Quantity	Positive if	Negative if
Object location (s)	Object is in front of surface	Object is in back of surface (virtual object)
Image location (s')	Image is in front of surface (real image)	Image is in back of surface (virtual image)
Image height (h')	Image is upright	Image is inverted
Radius (R_1 and R_2)	Center of curvature is in front of surface	Center of curvature is in back of surface
Focal length (f)	Concave mirror	Convex mirror

Where the front is, as usual, the part of the mirror that receives the light first from the object.



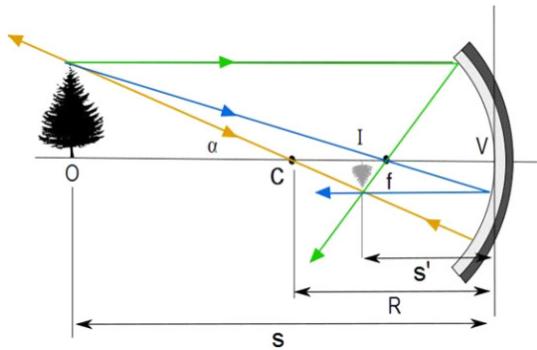
Notice that s' is negative for virtual images as always.

Ray Diagrams for Mirrors

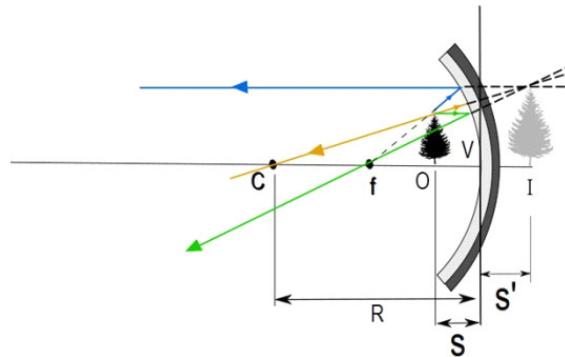
We have been drawing diagrams to find where images are formed for lenses, we should do the same for mirrors. We use a similar set of three rays. These rays are defined as follows:

Principal rays for a concave mirror:

1. Ray 1 is drawn from the top of the object such that its reflected ray must pass through f .
2. Ray 2 is drawn from the top of the object through the focal point to reflect parallel to the principal axis.
3. Ray 3 is drawn from the top of the object through the center of curvature. This ray will be incident on the mirror surface at a right angle and will be reflected back on itself.



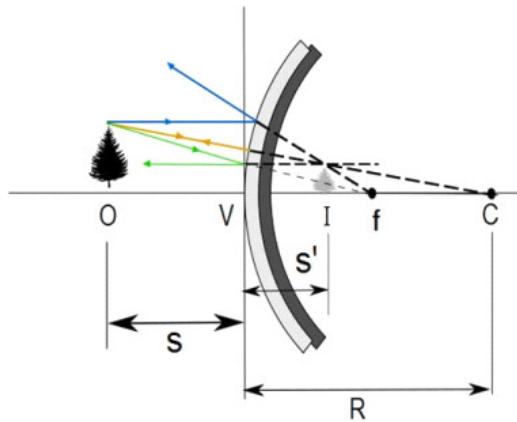
We can do the same for an object closer than a focal length



We also may have a mirror that curves, but curves the other way.

Principal rays for a convex mirror:

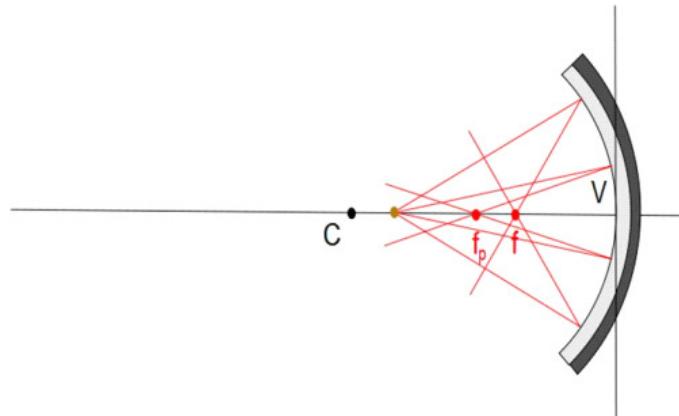
1. Ray 1 is drawn from the top of the object such that its reflected ray appears to have come from f .
2. Ray 2 is drawn from the top of the object to reflect parallel to the principal axis.
3. Ray 3 is drawn from the top of the object so that it appears to have come from the center of curvature. This ray will be incident on the mirror surface at a right angle and will be reflected back on itself.



Spherical Aberration

Question 223.16.12

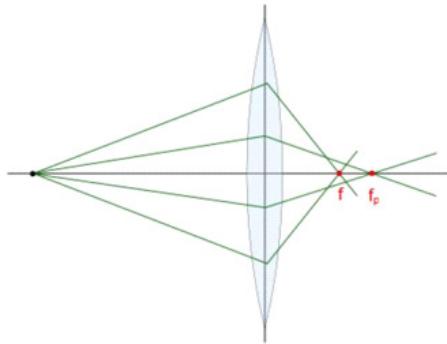
Spherical shapes are easier to make than parabolas or hyperbole, or other shapes. So optics manufacturers have been using spherical optics for centuries.



If we let rays converge from any direction from our spherical mirror we find we have a problem. The rays do not form a single image. Instead, they converge on a volume near where the image should be. Rays from larger angles converge at different distances than rays from small angles. This problem is known as *spherical aberration*. Most of the time, we will point our optics so the object is near the principal axis, so we can make the paraxial approximation that fixes this problem.

Question 223.16.5

The same problem happens with lenses



This problem is called spherical aberration and it was made famous as the main problem with the Hubble Telescope.

There are many aberrations that come from making lenses that are easy to manufacture, but that are not the perfect shape. We won't study these in this class. If you are curious, we cover these in PH375.

Just a note, we have run into another aberration, chromatic aberration, before. Mirrors in optical systems don't experience chromatic aberration. This is because mirrors in optical systems don't include a glass layer in front of the reflective surface like mirrors in your bathroom do. That glass is to protect the reflective surface from damage due to water (or toothpaste, etc.). In an optical system, this glass layer would cause unwanted reflection and absorption of the light, so it is not included. So mirrors don't have any refraction associated with them. This means that there will be no dispersion from a mirror.

17 Optical systems

Combinations of lenses

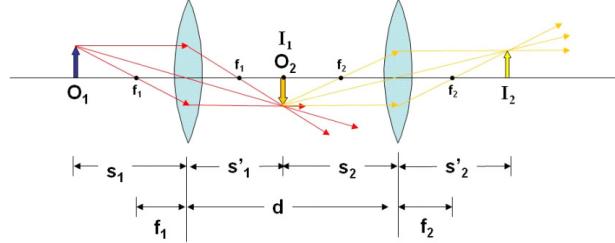
So far we have only used one lens or mirror at a time. But most optical systems are made from several lenses or mirrors (or a combination of both lenses and mirrors). We should think about how lenses work together to form optical systems like telescopes or microscopes or even compound camera lenses.

Question 223.17.1

Question 223.17.2

Question 223.17.3

To combine lenses, we do the same thing we did for the two surfaces of a thin lens. We form the image from the first lens as though the second lens is not there. Then we use the image from the first lens as the object for the second lens. Suppose we take two lenses of focal lengths f_1 and f_2 and place them a distance d apart.



Because this system would use a magnified image as the object for lens 2, the final magnification is the product of the two lens magnifications

$$M_{\text{combined}} = M_1 M_2 \quad (17.1)$$

Let's see that this must be true

$$M_1 = -\frac{s'_1}{s_1} = \frac{h'_1}{h}$$

and

$$M_2 = -\frac{s'_2}{s_2} = \frac{h'_2}{h'_1}$$

then

$$M_1 M_2 = \frac{h'_1}{h} \frac{h'_2}{h'_1} = \frac{h'_2}{h}$$

which is what we mean when we give the magnification of the optical system. We compare the output image size with the original object size.

It is a little more complicated to show where the final image will be. For the first lens we have

$$\frac{1}{s_1} + \frac{1}{s'_1} = \frac{1}{f_1} \quad (17.2)$$

where s'_1 is our first lens image distance. We can solve for s'_1

$$s'_1 = \frac{s_1 f_1}{s_1 - f_1} \quad (17.3)$$

We then take as the second object distance

$$s_2 = d - s'_1$$

we use the lens formula again.

$$\frac{1}{s_2} + \frac{1}{s'_2} = \frac{1}{f_2}$$

and again find the image distance

$$s'_2 = \frac{s_2 f_2}{s_2 - f_2}$$

but we can use our value of s_2 to find

$$\begin{aligned} s'_2 &= \frac{(d - s'_1) f_2}{(d - s'_1) - f_2} \\ &= \frac{(d - s'_1) f_2}{d - s'_1 - f_2} \end{aligned}$$

We have an expression relating the image distances, d and f_2 . But we would really like to have an expression that relates s_1 and s'_2 . Lets use

$$s'_1 = \frac{s_1 f_1}{s_1 - f_1}$$

and substitute it into our expression for s'_2

$$s'_2 = \frac{\left(d - \frac{s_1 f_1}{s_1 - f_1}\right) f_2}{d - \frac{s_1 f_1}{s_1 - f_1} - f_2}$$

This looks messy, but we can do some simplification

$$s'_2 = \frac{df_2 - \frac{s_1 f_1 f_2}{s_1 - f_1}}{d - f_2 - \frac{s_1 f_1}{s_1 - f_1}} \quad (17.4)$$

Well, it is still a little messy, but we have achieved our goal. We have s'_2 in terms of the focal lengths, d , and s_1 .

Suppose we let $d \rightarrow 0$. Then

$$\begin{aligned}s'_2 &= \frac{-\frac{s_1 f_1 f_2}{s_1 - f_1}}{-f_2 - \frac{s_1 f_1}{s_1 - f_1}} \\&= \frac{\frac{s_1 f_1 f_2}{s_1 - f_1}}{\frac{f_2(s_1 - f_1)}{s_1 - f_1} + \frac{s_1 f_1}{s_1 - f_1}} \\&= \frac{s_1 f_1 f_2}{f_2 s_1 - f_2 f_1 + s_1 f_1} \\&= \frac{s_1 f_1 f_2}{s_1 (f_2 + f_1) - f_2 f_1}\end{aligned}$$

So

$$s'_2 = \frac{s_1 f_1 f_2}{s_1 (f_2 + f_1) - f_2 f_1}$$

Lets undo the math that brought us s'_2 in the first place

$$\begin{aligned}\frac{1}{s'_2} &= \frac{s_1 (f_2 + f_1) - f_2 f_1}{s_1 f_1 f_2} \\&= \frac{s_1 (f_2 + f_1)}{s_1 f_1 f_2} - \frac{f_2 f_1}{s_1 f_1 f_2} \\&= \frac{(f_2 + f_1)}{f_1 f_2} - \frac{1}{s_1}\end{aligned}$$

or

$$\frac{1}{s'_2} + \frac{1}{s_1} = \frac{(f_2 + f_1)}{f_1 f_2}$$

Which looks very like the lens formula with

$$\frac{1}{f} = \frac{(f_2 + f_1)}{f_1 f_2}$$

If we unwind this expression, we find

$$\begin{aligned}\frac{1}{f} &= \frac{f_2}{f_1 f_2} + \frac{f_1}{f_1 f_2} \\&= \frac{1}{f_1} + \frac{1}{f_2}\end{aligned}\tag{17.5}$$

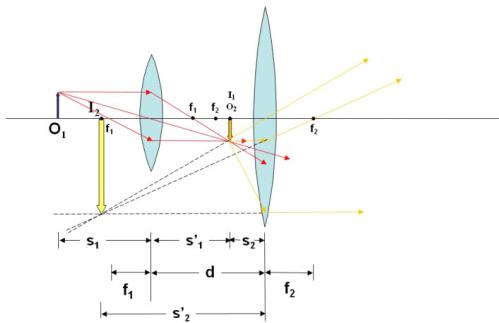
This is how we combine thin lenses. We see that the two lenses are equivalent to a single lens with focal length f as long as they are close together.

Of course, we had to place our lenses right next to each other for this to work. This is not the case for a telescope or microscope. We should look at such a case. There is no need for more math. We can go back to equation (17.4).

$$s'_2 = \frac{df_2 - \frac{s_1 f_1 f_2}{s_1 - f_1}}{d - f_2 - \frac{s_1 f_1}{s_1 - f_1}}$$

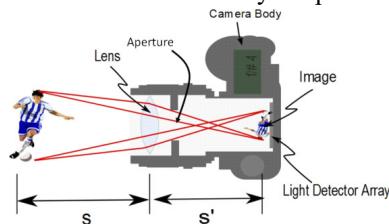
But let's look at a case using ray diagrams. For this case, let's take two lenses, and let's have the first lens make a real image. Once again, let's have that image be the object

for the second lens. But this time, let's move the second lens so that the image from the first lens (object for the second lens) is closer to the second lens than f_2 . If that is the case, the second lens works like a magnifier. The final image is enlarged.



The Camera

in 1900 George Eastman introduced the Brownie Camera. This event has changed society dramatically. The idea behind a camera is very simple.



The camera has a lens (often a compound lens like the ones we have just discussed) and a screen for projecting a real image created by the lens.

Let's take an example camera. Say we wish to take a picture of Aunt Sally. Aunt Sally is about 1.5 m tall. She is standing about 5 m away. Then to fit the image of Aunt Sally on our 35 mm detector, we must have

$$\begin{aligned} h &= 1.5 \text{ m} \\ h' &= 0.035 \text{ m} \\ s &= 5 \text{ m} \\ f &= 0.058 \text{ m} \end{aligned}$$

We wish to find s' and m . Let's do m first.

$$\begin{aligned} m &= \frac{h'}{h} = \frac{-0.035 \text{ m}}{1.5 \text{ m}} \\ &= -2.333 \times 10^{-2} \end{aligned}$$

so our image is small and inverted. The small size we wanted. But now we know that the image in our cameras is upside down. A digital camera uses its built-in computer to turn the image right side up for us on the display on the back of the camera.

Now let's find s' .

$$\begin{aligned} s' &= \frac{fs}{s-f} \\ &= 5.8681 \times 10^{-2} \text{ m} \\ &= 58.681 \text{ mm} \end{aligned}$$

so our detector must be 58.681 mm from the lens.

Now suppose we want to photograph a 1000 m tower from 2 km away. Then

$$\begin{aligned} m &= -\frac{0.035 \text{ m}}{1000 \text{ m}} \\ &= -3.5 \times 10^{-5} \end{aligned}$$

and

$$\begin{aligned} s' &= \frac{(0.058 \text{ m})(2000 \text{ m})}{2000 \text{ m} - (0.058 \text{ m})} \\ &= 5.8002 \times 10^{-2} \text{ m} \\ &= 58.002 \text{ mm} \end{aligned}$$

Notice that the image distance changed, but not by very much. This is why you need a focus adjustment on the lens of a good camera. Objects far away require a different s' value than objects that are close. Usually you twist the lens housing to make this adjustment. The lens housing has a threaded screw system that increases or decreases s' as you twist. Cell phones and consumer cameras often have a motor that makes this adjustment for you. In some cameras you may see the lens move back and forth as someone takes a picture.

There are several things that govern whether a picture will be good. When you buy a quality manual lens, it will be marked in $f/\#s$. The specification of an automatic lens will be given in terms of $f/\#s$. To help us buy such lenses, we should understand what the terminology means.

Question 223.17.4

Most things we want to take a snapshot of are much farther than 58 mm from the

camera. For such objects we can revisit the magnification.

$$m = -\frac{s'}{s}$$

but from the thin lens formula

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

Question 223.17.5 If $s \gg f$ then we can say that $1/s \approx 0$ and so $s' \approx f$. Then

$$m = -\frac{f}{s}$$

and we see that the size of the image is directly proportional to the focal distance. If we change the focal distance, we can change the size of the image. This is how a zoom lens works. A zoom lens is a compound lens, and the focal length is changed by increasing the distance between the component lenses. This is what your camera is doing when it zooms in and out when you push the telephoto button.

Remember we studied intensity

$$I = \frac{P}{A}$$

Photographic film and digital focal plane arrays detect the intensity of light falling on them. We can see that the area of our image depends on our magnification, which depends on s' and for our distant objects it is proportional to f . The image area is proportional to $s'^2 \approx f^2$. So we can say that the area is proportional to f^2 . Then

$$I \propto \frac{P}{f^2}$$

Question 223.17.6

The power entering the camera is proportional to the size of the aperture (hole the light goes through). A bigger aperture lets in more light. A smaller aperture lets in less light. If the camera has a circular opening, this area is proportional to the square of the diameter of the opening, D^2 so

$$I \propto \frac{D^2}{f^2}$$

This ratio is useful because it tells us how much intensity we get in terms of things we can easily know. Good cameras have changeable aperture sizes, and good lenses have changeable focal lengths. By using the combination of these two terms, we can ensure we will get enough light (but not too much) when we take the picture.

It would be good to give this ratio a special name. But instead, we named the ratio

$$f/\# \equiv \frac{f}{D} \quad (17.6)$$

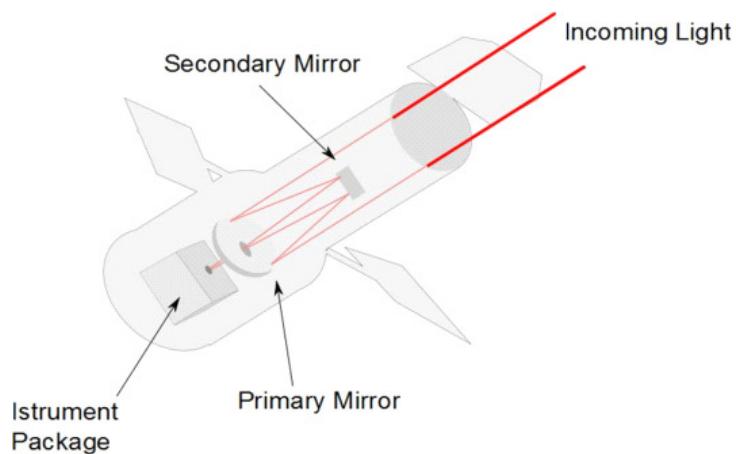
It is called the $f/\#$ (pronounced f-number) so

$$I \propto \frac{1}{(f/\#)^2} \quad (17.7)$$

So good cameras have adjustable lens systems marked in $f/\#$'s. Typical values are

$f/2.8$, $f/4$, $f/5.6$, $f/8$, $f/11$, and $f/16$.

This terminology is used for telescope design as well. The Hubble telescope is an $f/24$ Ritchey-Chretien Cassegrainian system with a 2.4 m diameter aperture. The effective focal length is 57.6 m.



It is important to realize that electronic (and biological) sensors don't react instantly to what we see. The intensity is

$$I = \frac{P}{A} = \frac{\Delta E}{\Delta t A}$$

so there is a time involved. The time it takes to collect enough light to form an image on the sensor is called the *exposure time*.

$$\Delta t = \frac{\Delta E}{IA}$$

So changing our $f/\#$ changes the needed exposure time by changing the intensity. How sensitive our camera sensor is also affects the exposure time. Modern sensors have adjustable sensitivity. The photography world gives the three letters ISO for the name for this detector sensitivity. There isn't a standard for exactly what ISO setting gives what exposure. Different manufacturers use slightly different numbers. But a change in ISO settings usually are equivalent to one $f/\#$ change in exposure.

This is part of what a good photographer does in taking a picture. The photographer will adjust the $f/\#$ and the exposure time and ISO to get a photograph that is not too exposed (too light) or underexposed (too dark).

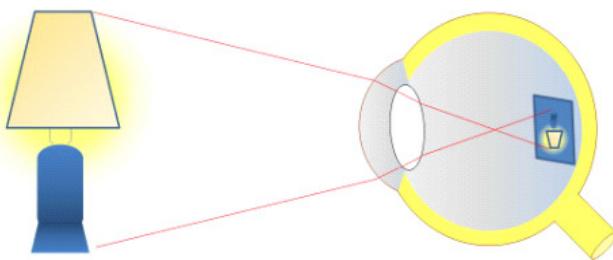
18 Eyes and magnifiers

Fundamental Concepts

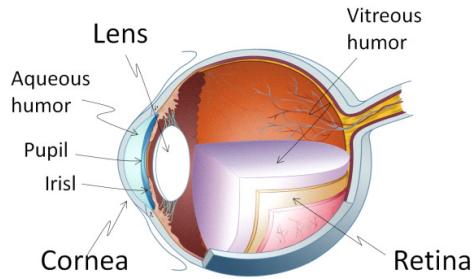
- Angular magnification compares the apparent size of an image with and without an optical system.
- The power of a lens is measured in Diopters which are defined to be $1/f$ (m)
- Compound magnifiers use an objective lens to form an image, and an eyepiece to magnify the image.

The Eye

Question 223.18.1

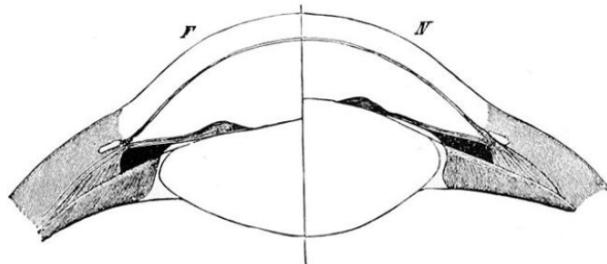


The figure above shows the parts of the eye. The eye is like a camera in its operation, but is much more complex. It is truly a marvel. The parts that concern us in this class are the cornea, crystalline lens, pupil, and the retina.



Question 223.18.2

The Cornea-lens system refracts the light onto the retina, which detects the light. The lens is focused by a set of muscles that flatten the lens to change its focal length. The focusing process is different from a standard camera. The camera moves the lens to achieve a different image distance. Our eye can't change the distance between the lens system and the retina. So our eye changes the shape of the lens, changing its focal length.



The crystalline lens becomes thicker, and therefore more curved when the ciliary muscle flexes. Austin Flint, "The Eye as an Optical Instrument," *Popular Science Monthly*, Vol. 45, p203, 1894 (Image in the public domain)

The focusing system is called accommodation. This system becomes less effective at about 40 because the lens becomes less flexible. The closest point that can be focused by accommodation is called the near point. It is about 25 cm on average. There is, of course, no such thing as an average person. All of us are a little bit different. You young students probably have a much shorter near point than 25 cm. For those of us that are a little older, 25 cm or more is more likely.

The farthest point that can be focused is ideally a long way away. It is called the far point. Both the near and far points degrade with years leading to bifocals.

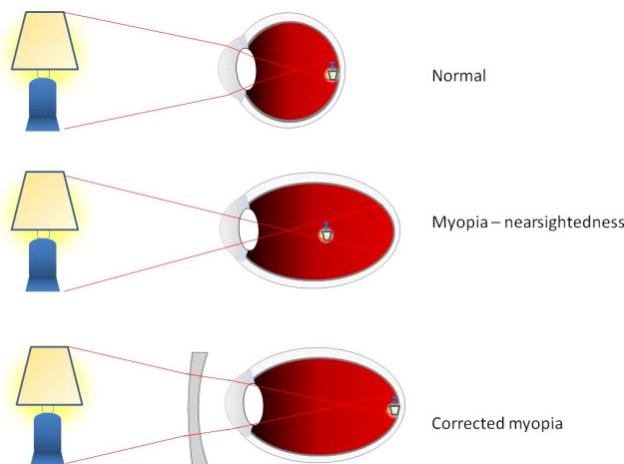
The iris changes the area of the pupil (the aperture of the eye). The pupil is, on average,

about 7 mm in diameter.

Nearsightedness

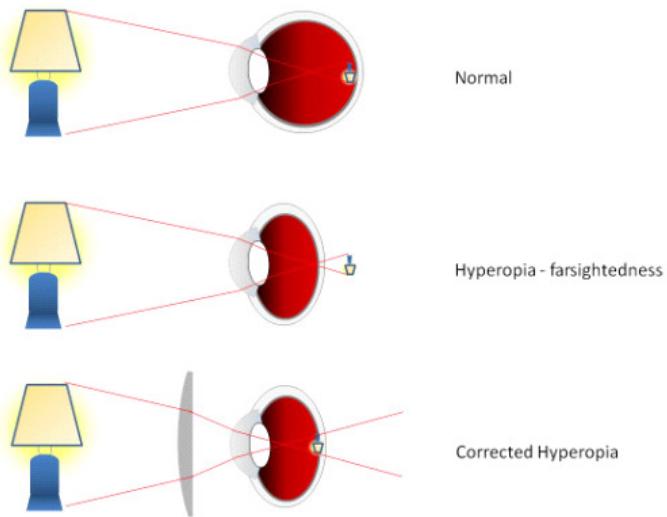
Question 223.18.3

In some people the cornea-lens system focuses in front of the retina. This is called nearsightedness or myopia.



Farsightedness

Sometimes the cornea-lens system focuses in back of the retina. This is called farsightedness or hyperopia. It is corrected with a converging lens



Diopters

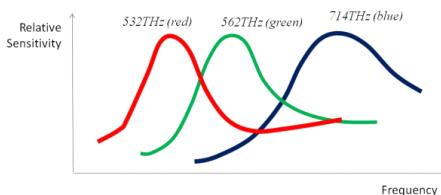
Question 223.18.4

Eye glasses use a different unit of measure to describe how they bend light. The unit is the *diopter* and it is equal one over the focal length

$$\text{diopter} = \frac{1}{f \text{ (m)}} \quad (18.1)$$

Color Perception

The eye detects different colors. The receptors called cones can detect red, green, and blue light.



The eye combines the red, green, and blue response to allow us to perceive many different colors.

Most digital cameras also have red, green, and blue pixels to provide color to images. The detectors in digital cameras are often have much narrower frequency bands than the eye. Likewise, television displays and monitors have red, green, and blue pixels.

By targeting the eye receptors, power need not be wasted in creating light that is not detected well by the eye. The difference in band width can cause problems in color mixing. Yellow school busses (perceived as different amounts of green and red light) may be reddish or green if the bandwidths are chosen poorly.

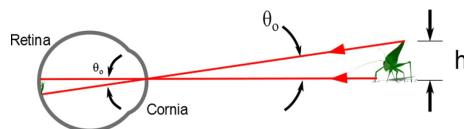
The science of human visual perception of imagery is called *image science*. There are many applications for this field, from forensics to intelligence gathering.

Optical Systems that Magnify

Simple Magnifier

You may have noticed that, so far, when we say “magnification” we are defining it in a way that is different than every-day usage of the word. We defined magnification to be how big the image is compared to how big the object is. But in every-day speech, magnification means how big the image is using a lens or optical system compared to how big it looks without the lens or optical system. We will call this kind of magnification the *angular magnification*.

We already encountered the simple magnifier when we studied ray diagrams. Let's use the simple magnifier to define angular magnification. To understand angular magnification, we can use what we know about easy rays to draw the rays that go straight through the lens of the eye. If we pick a ray from the top of our object that goes through the center of the lens, that ray won't seem to change direction at all. It will hit the retina to form the top of the image of the object. We can do the same for the bottom of the object. Then we can see from the next figure



that the angle θ_o subtends both the object and the image of the object. If the angle increases, so does the size of the image on the retina.

If we move the object closer, θ_o increases, and so does the size of the image. When we get to about 25 cm, we reach the limit of the eye for focusing. If we move the object any closer, it will appear fuzzy. We called this position, the closest point where we can place an object and still bring it into focus with our eye, the *near point*. Thus the

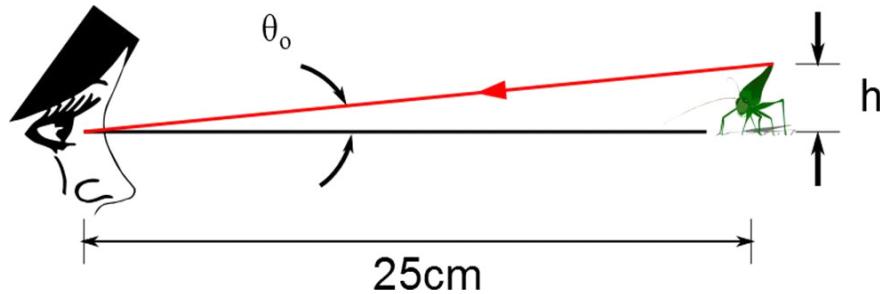


Figure 18.12.

maximum value of θ_o will be at the near point for unaided viewing.

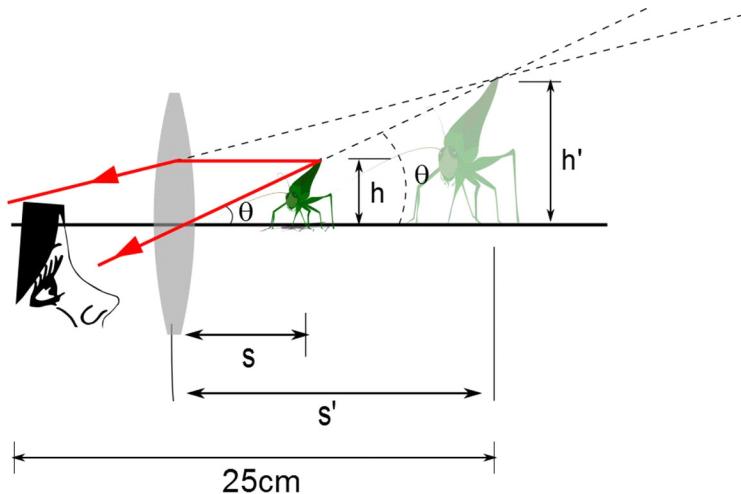


Figure 18.13.

But suppose we want to see this object in more detail. We can use a magnifying glass. If we place the object closer to the magnifying glass than the focal distance ($s < f$), then (lower part of the figure) we have a virtual image with magnification

$$m = \frac{-s'}{s} \quad (18.2)$$

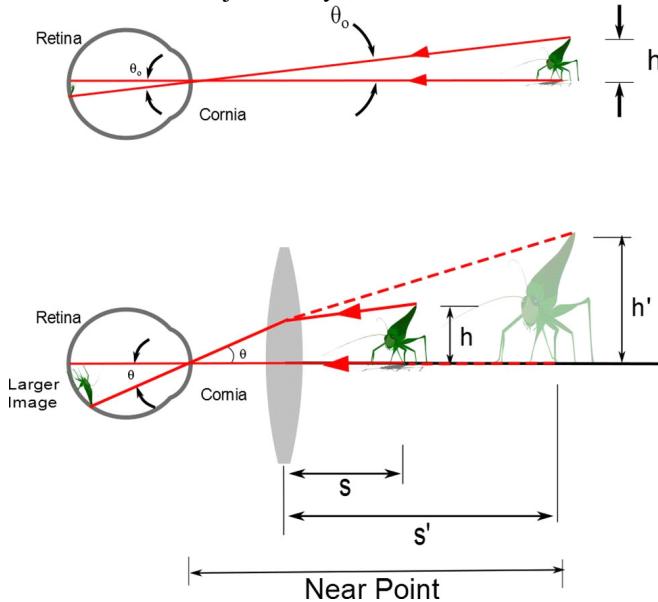
which is larger than one and positive (because s' is negative).

But what we really want to know is how much bigger the image looks with the lens than it did without the lens. We define the angular magnification

$$M = \frac{\theta}{\theta_o} \quad (18.3)$$

This is the ratio of the image sizes with and without the magnifier lens.

This is really different than the magnification we have studied before. The magnification we have been using compared the size of the image with the size of the object. So, the angular magnification compares how big the object seems to be with and without a lens or lens system. We can think of this as a comparison between the size of the real image on the retina formed with just our eye, and the one formed with the magnifier.



If the virtual image formed is farther than the near point of the eye, ($s' > \sim 25 \text{ cm}$) it will seem smaller than it would be at the near point because it is farther away. If the virtual image is closer than the near point, it will be fuzzy because the eye cannot focus closer than the near point. Thus, the value of M will be maximum when s' is at the near point of the eye. We can find where to place the image so that we get maximum magnification. Taking just the magnifier,

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

$$\frac{1}{s} + \frac{1}{-25 \text{ cm}} = \frac{1}{f}$$

and

$$\frac{1}{s} = \frac{-25 \text{ cm} - f}{-f(25 \text{ cm})}$$

or

$$s = \frac{(25 \text{ cm}) f}{25 \text{ cm} + f} \quad (18.4)$$

Returning to figure (18.13). Note that the person has adjusted her viewpoint so the ray that passes through the middle of the lens also passes through the middle of her eye lens (cornea). So the angle θ in the figure is also the angle that subtends the image on her retina. This was nice of her because it makes our math easier. Using small angle approximations, we can write

$$\tan \theta_o = \frac{h}{25 \text{ cm}} \approx \theta_o$$

and

$$\tan \theta = \frac{h}{s} \approx \theta$$

then the maximum angular magnification is

$$\begin{aligned} m_{\max} &= \frac{\theta}{\theta_o} = \frac{\frac{h}{s}}{\frac{h}{25 \text{ cm}}} \\ &= \frac{25 \text{ cm}}{\frac{25 \text{ cm}f}{25 \text{ cm}+f}} \\ &= \frac{25 \text{ cm} + f}{f} \\ &= 1 + \frac{25 \text{ cm}}{f} \end{aligned}$$

We can also find the minimum magnification by letting s be at f . This gives

$$\theta = \frac{h}{f}$$

$$\begin{aligned} m_{\min} &= \frac{\theta}{\theta_o} = \frac{\frac{h}{f}}{\frac{h}{25 \text{ cm}}} \\ &= \frac{25 \text{ cm}}{f} \end{aligned}$$

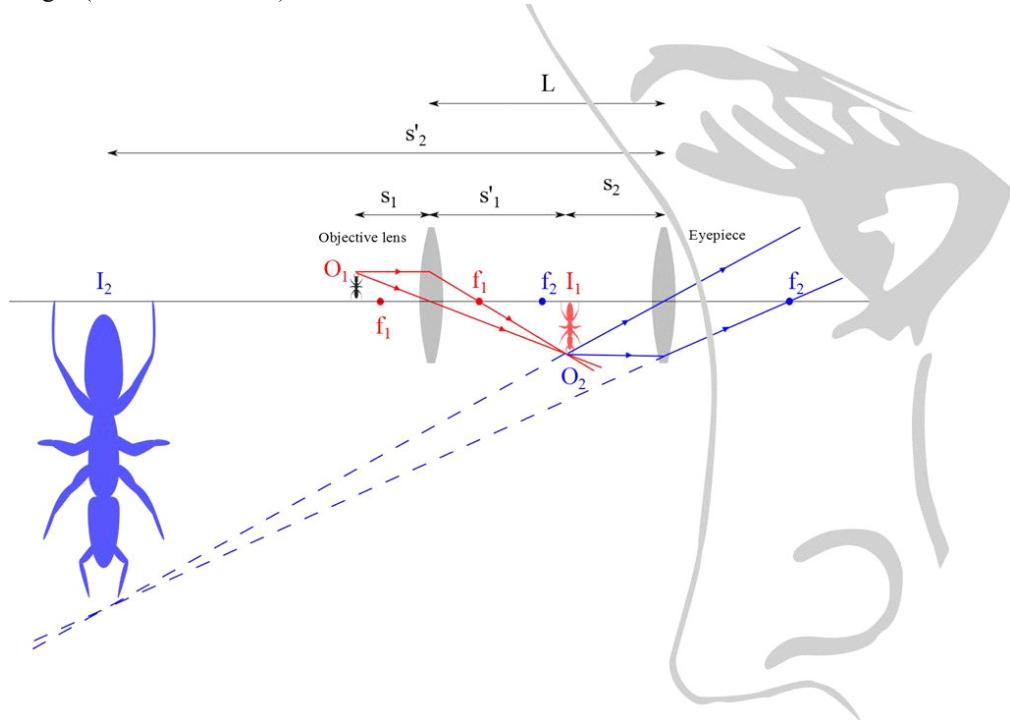
When you use a magnifying glass, notice that you move the lens back and forth between these extremes until you can see the level of detail you want.

But the idea of a magnifier is more than just seeing the details small objects. We use the idea of a simple magnifier in combination with other lenses to make the magnification happen in telescopes, microscopes, and other instruments that magnify.

The Microscope

To see things that are very small, we add another lens to our simple magnifier. We will place this lens near the object. Since this new lens is near the object, let's give it the name *objective lens* or just *objective*. We will keep a simple magnifier and place it near the eye. Since our simple magnifier is near our eye, let's call it the *eyepiece*.

The objective will have a very short focal length. The eyepiece will have a longer focal length (a few centimeters).



We separate the lenses by a distance L where

$$L > f_o$$

$$L > f_e$$

We place the object just outside the focal point of the objective. The image formed by the objective lens is then real and inverted. We use this image as the object for the eyepiece. The image formed is upright and virtual, but it looks upside down because the object for the eyepiece (first image for the objective) is upside down.

The magnification of the first lens is

$$m_o = \frac{-s'_1}{s_1} \approx -\frac{L}{f_1}$$

because $s_1 \approx f_1$ and $s'_1 \approx L$. The magnification of the eyepiece is just that of a simple magnifier when the object is placed at the focal point f_1

$$m_e = \frac{-s'_2}{s_2} \approx \frac{25 \text{ cm}}{f_2}$$

The combined magnification is

$$m = m_o m_e = -\frac{L}{f_1} \frac{25 \text{ cm}}{f_2} \quad (18.5)$$

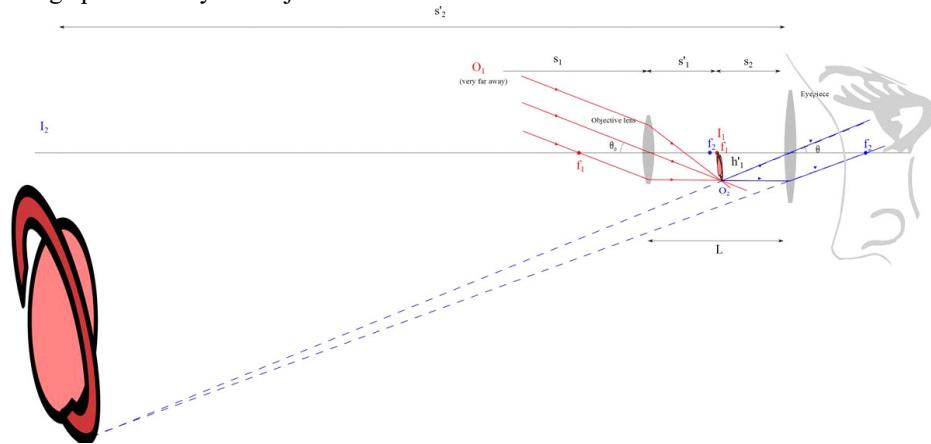
this is the minimum magnification.

Telescopes

There are two types of telescopes *refracting* and *reflecting*. We will study refracting telescopes first.

Refracting Telescopes

Like the microscope, we combine two lenses and call one the objective and the other the eyepiece. The eyepiece again plays the role of a simple magnifier, magnifying the image produced by the objective.



We again form a real, inverted image with the objective. We are now looking at distant objects, so the image distance $s'_o \approx f_o$. We use the image from the objective as the object for the eyepiece. The eye piece forms an upright virtual image (that looks inverted because the object for the eyepiece is the image from the objective, and the real image from the objective is inverted). The largest magnification is when the rays exit the eyepiece parallel to the principal axis. Then the image from the eyepiece is formed at infinity (but it is very big, so it is easy to see). This gives a lens separation of $f_o + f_e$ which will be roughly the length of the telescope tube.

The angular magnification will be

$$M = \frac{\theta}{\theta_o} \quad (18.6)$$

where θ_o is the angle subtended by the object at the objective. That is the angle we would have with no lenses and just our eye, because it is the angle subtended by the object without the optical system. The angle θ is subtended by the final image at the viewer's eye using the optical system. Consider s_o is very large. We see from the figure that

$$\tan \theta_o = -\frac{h'}{f_o} \quad (18.7)$$

and with s_o large we can use small angles.

$$\theta_o = -\frac{h'}{f_o} \quad (18.8)$$

The angle θ will be the angle formed by rays bent by the lens of the eye. This angle will be the same as the angle formed by a ray traveling from the tip of the first image and traveling parallel to the principal axis. This ray is bent by the objective to pass through f_e . Then

$$\tan \theta = \frac{h'}{f_e} \approx \theta \quad (18.9)$$

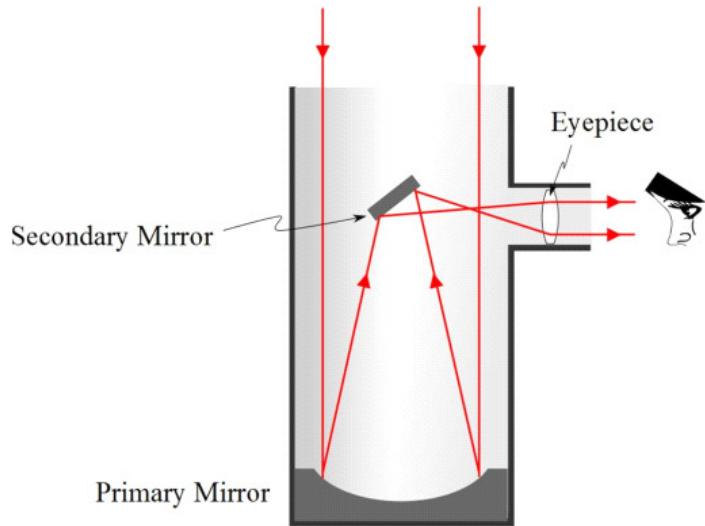
so

The magnification is then

$$m = \frac{\theta}{\theta_o} = \frac{\frac{h'}{f_e}}{-\frac{h'}{f_o}} = -\frac{f_o}{f_e} \quad (18.10)$$

Reflecting Telescopes

Reflecting telescopes use a series of mirrors to replace the objective lens. Usually, there is an eyepiece that is refractive (although there need not be, radio frequency telescopes rarely have refractive pieces).



There are two reasons to build reflective telescopes. The first is that reflective optics do not suffer from chromatic aberration. The second is that large mirrors are much easier to make and mount than refractive optics. The Keck Observatory in Hawaii has a 10 m reflective system. The largest refractive system is a 1 m system. The Hubble telescope has a 2.5 m aperture.

The telescope pictured in the figure is a Newtonian, named after Newton, who designed this focus mechanism. Many other designs exist. Popular designs for space applications include the Cassegrain telescope.

The rough design of a reflective telescope can be worked out using refractive pieces, then the rough details of the reflective optics can be formed.

19 Resolution and Charge

Fundamental Concepts

- Two points can be distinguished when imaged if their angular separation is a minimum of $\theta_{\min} = 1.22 \frac{\lambda}{D}$
- There is a property of matter called “charge.”
- There seem to be two types of charges, called “positive” and “negative.”

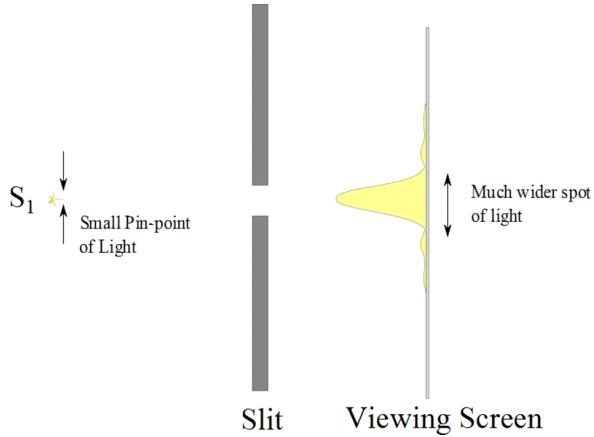
Resolution

Question 223.19.1

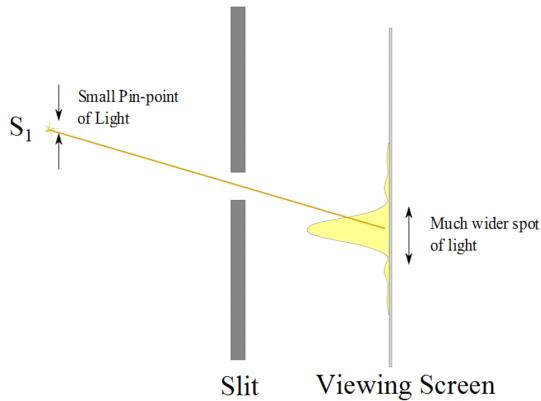
We have emphasized that an extended object can be viewed as a collection of point objects. Then the image is formed from the collection of images of those point objects. It would be great if optical systems could form images with infinite precision—that is, the image of a point object would be a point image. The fact that light acts as a wave prevents this from being true. The quality of our image depends on how poorly a point object is imaged. If each point object makes a large circle of light on the screen or detector array, we get a very confusing image (it will look blurry to us).¹⁴ Let’s see why this will happen so we can know how to minimize the effect.

We already know that if we take light and pass it through a single slit, we get an intensity pattern that has a central bright region.

¹⁴ In Fourier Optics, the intensity pattern that comes from imaging a single point is called a *point spread function* because it shows how spread out the light from a single point will be. In mechanical engineering, we might call this an impulse response function. It is the same idea applied to optics.

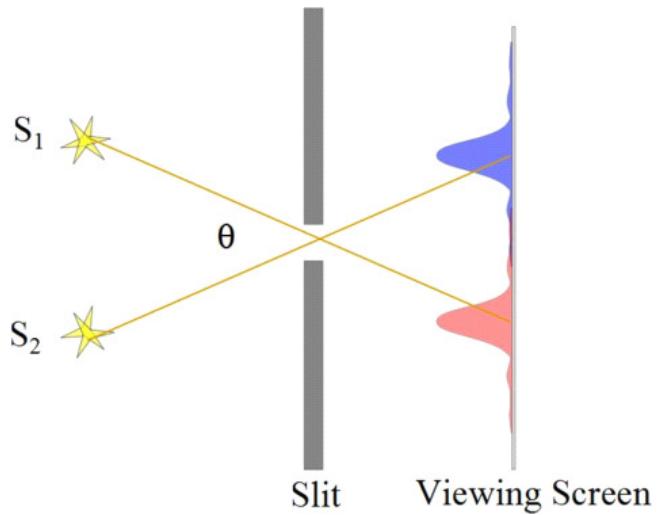


Remember that normal objects will be made up of many small points of light (either due to reflection or glowing) and each of these will form such an intensity pattern on a screen. Here is a bright point source that is not on the axis, and we see that it too makes a bright spot on the screen (and smaller bright spots or rings, depending on the shape of the aperture)

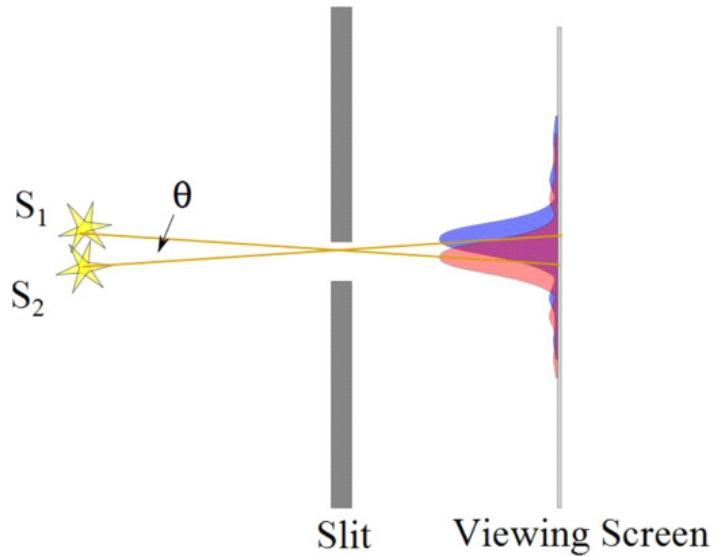


So our images will be made up of many central bright spots, each of which represents a point of light from the object. These central bright spots may overlap, (and their secondary maxima certainly will overlap).

Let's take a simple case of two points of light, S_1 and S_2 . If we take a single slit and pass light from two distant point sources through the slit, we do not get two sharp images of the point sources. Instead, we get two diffraction patterns.



If these patterns are formed sufficiently far from each other, it is easy to tell they were formed from two distinct objects. Each point became a small blur, but that is really not so bad. We can still tell that the two blurs came from different sources. If our pixel size is about the same size of the blur, we won't even notice the blurriness in the digital imagery.



But if the patterns are formed close to each other, it gets hard to tell whether they were formed from two objects or one bright object. We now have a problem. Suppose you are trying to look at a star and see if it has a planet. But all you can see is a blur. You

can't tell if there is one source of light or two.

Long ago an early researcher titled Lord Rayleigh developed a test to determine if you can distinguish between two diffraction patterns.

When the central maximum of one point's image falls on the first minimum of another point's image, the images are said to be just resolved.

This test is known as *Rayleigh's criterion*.

We can find the required separation for a slit. Remember that

$$\sin(\theta) = m \frac{\lambda}{a} \quad m = \pm 1, \pm 2, \pm 3 \dots \quad (19.1)$$

gives the minima. We want the first minimum, so

$$\sin(\theta) = \frac{\lambda}{a} \quad (19.2)$$

If we place the second image maximum so it is just at this location, the two images will be just barely resolvable. In the small angle approximation, $\sin(\theta) \approx \theta$ so

$$\theta_{\min} = \frac{\lambda}{a} \quad (19.3)$$

Now you may be saying to yourself that you don't often take pictures through single illuminated slits, so this is nice, but not really very interesting.

Suppose, instead, that we image a circular aperture. Again, we won't go through all the math (there are Bessel functions involved) but the criterion becomes

$$\theta_{\min} = 1.22 \frac{\lambda}{D} \quad (19.4)$$

where D is the aperture diameter.

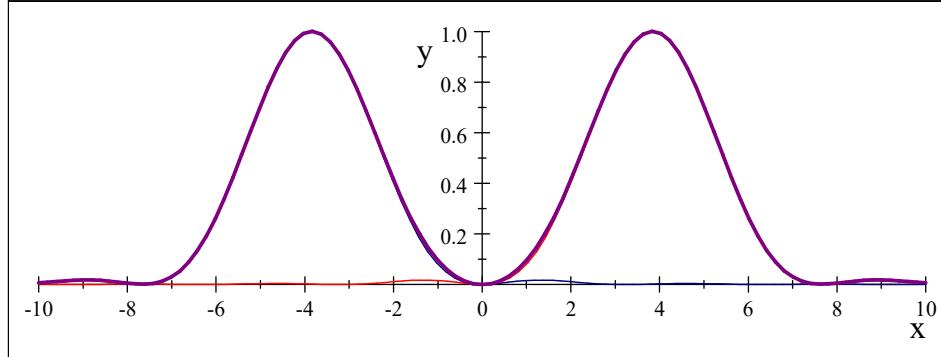
Still, you may say, I don't like pictures taken through small circles any better than through small slits! Yet, in fact, you do. Most cameras have circular apertures. The light that passes into your phone camera must pass through the circular lens. For that matter, the pupil of our eye is a circular aperture. So most images we see are made using circular apertures.

The Rayleigh criteria tells you, based on your camera aperture size, how a point source will be imaged on the film or sensor array. If we consider extended sources (like your favorite car or Aunt Matilda) to be collections of many point sources, then we have a way to tell what features will be clearly resolved on the image and what features will not (like you may not be able to see the lettering on the car to tell what model it is, or you may not be able to distinguish between the gem stones in Aunt Matilda's necklace).

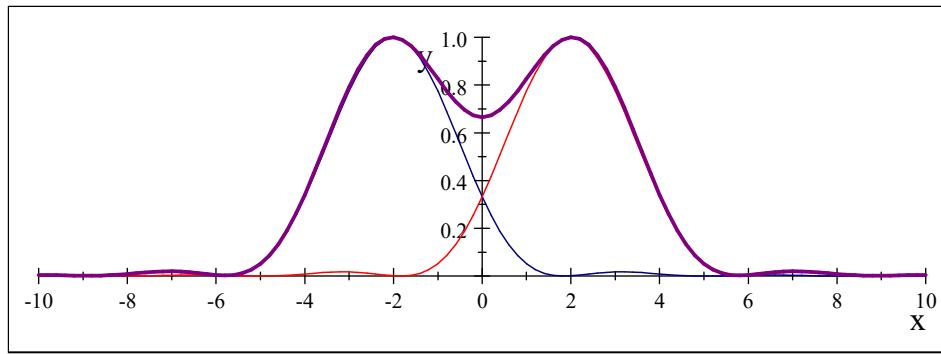
Question 223.19.2

because the image is too blurry to see these features clearly).

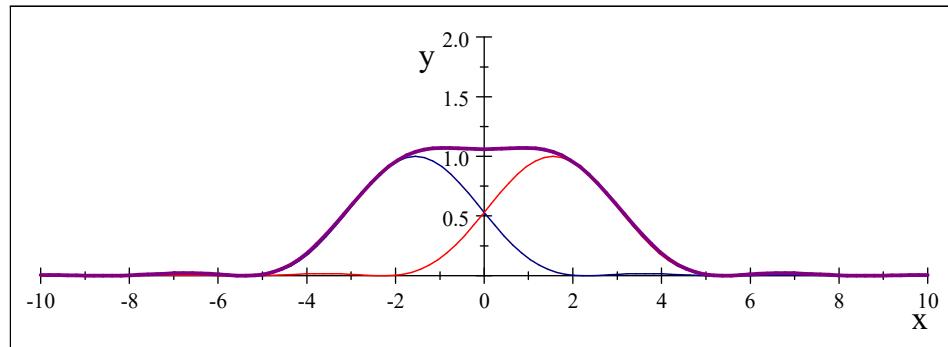
In the next three figures you can see two easily resolved intensity patterns, then two that can be resolved using Rayleigh's criterion, and finally a more modern astronomical resolution criterion made by a researcher called Sparrow. Finally there is figure with two intensity patterns that would not be resolvable (would look like just one point of light).



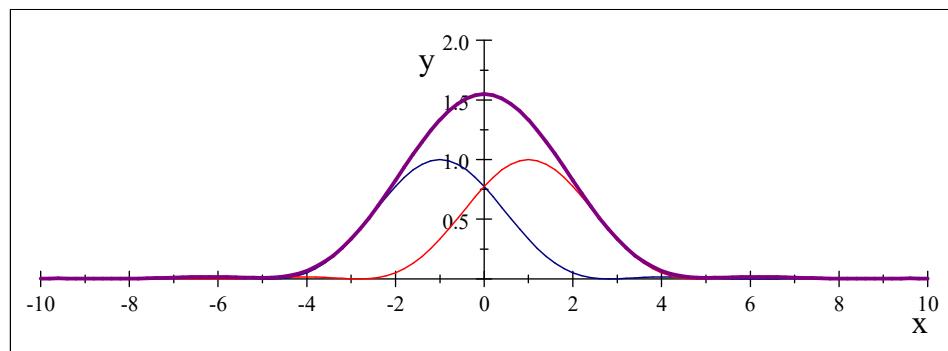
Pattern from two circular sources.



Rayleigh Criteria: Pattern from two circular sources where the sources are close enough that the maximum from one pattern is placed on the minimum of the other. Lord Rayleigh gave this as the criteria for just barely being resolved.

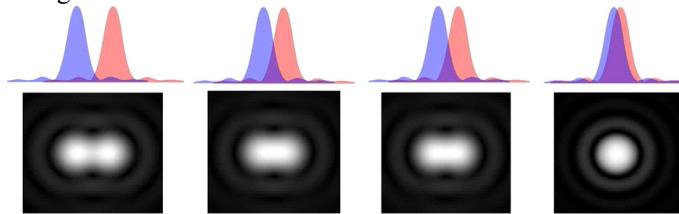


Sparrow Criteria: This is a less conservative resolution criteria than Rayleigh. When the intensity pattern is flat on the top, there must be two sources. This criterial is used in astronomy.



Two circular sources unresolved.

Here are our four cases again but with pictures that show what you would see in an image of the two light sources for each criterion.



From our equation we can see that we have better resolution if D is bigger. This is why professional photographers use large lenses and not their cell phones. The cell phone cameras have apertures that are a few millimeters. Typical professional cameras have 67 mm apertures. We can see that for a cell phone the angle for minimal resolution is

about

$$\theta_{\min} = 1.22 \frac{500 \text{ nm}}{3 \text{ mm}} = 2.0333 \times 10^{-4} \text{ rad}$$

For the professional lens, the minimum resolution is about

$$\theta_{\min} = 1.22 \frac{500 \text{ nm}}{67 \text{ mm}} = 9.1045 \times 10^{-6} \text{ rad}$$

That is a whopping factor of 22 better resolution. If you need to find a small crack in a structure, or if you want to print a wall sized portrait of your Aunt Miltilda, the extra resolution might be necessary for your application.

Charge model

So far we have claimed that light is a wave in an electromagnetic field. We talked about light waves being made in the electromagnetic field by moving charges. But we have not proved it this to be the case. We will find that it will take the rest of the semester to do so! We ned to start by looking at charge and what makes charges move.

But let's think about this conceptually and see if we can motivate our study.

Question 223.19.3

PHET Radio Wave Applet

We know that there is an electromagnetic spectrum, and that visible light is just a small part of that spectrum. Radio waves are also part of the spectrum of light.

And we should review, how are radio waves produced? We know electricity is involved.

The answer is that charged particles, like the electrons flowing through the antenna of our radio station, create an electromagnetic field. That field is drug along when the electrons move in the antenna. If we make the electrons oscillate, we can make waves in the field. This is much like having a 3rd grade class all hold the edges of a parachute and having the 3rd graders jump up and down. Waves are made in the parachute.

But what is charge? How do we know there are such things as charged particles?

That is the subject we will take up next. Then we will study the motion and actions of these charged particles. Finally we will show that the fields made by charged particles can act as a medium for waves, and that there is good evidence that those waves exist.

Evidence of Charge

Balloon and 2 by 4 demo

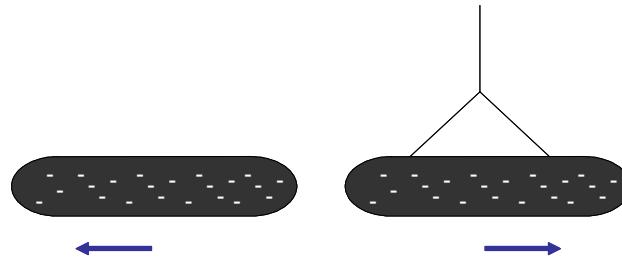
Balloon on wall demo

Comb and bits of paper demo

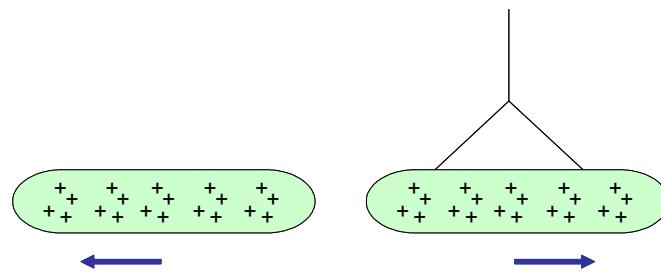
Let's start with something we all know. Let's rub a balloon in someone's hair. If we do this we will find that the balloon sticks to the wall. Why?

We say the balloon and comb have become *charged*. What does this mean? We will have to investigate this more as we learn more about how matter is structured, but for now let's assume charge is some property that provides this phenomena we have observed with the balloon (i.e. it sticks to the wall). Now lets try rubbing other things. We could rub two rubber or plastic rods.

Glass and Rubber Rod Demo

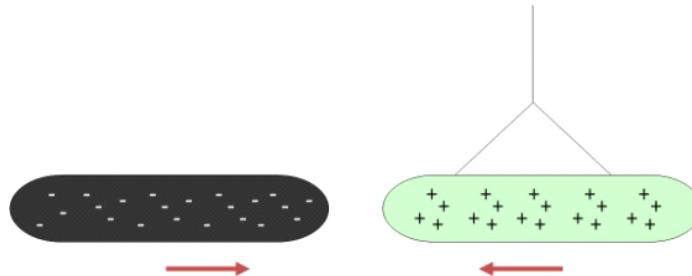


Two charged rubber rods are placed close together. The rods repel each other. and we could also rub two glass rods



Notice that in each case we have created a force between the two rods. The rods now repel each other.

Now let's try a glass and a rubber rod



Now the two different rods attract each other.

Notice that in our demo, rods that are the same repel and rods that are different attract.

We make the intellectual leap that the different rods have different charges. So we are really saying:

1. There are two types of charge.
2. Charges that are the same repel one another and charges that are different attract one another.
3. Friction seems to produce charge, but you have to rub the right materials together.

We will call the rubber or plastic rod charges *negative* and the glass rod charges *positive* but the choice is arbitrary. Ben Franklin is credited with making the choice of names. He really did not know much about charge, so he just picked two names (we will see that in some ways his choice was somewhat unfortunate, but hay, he was an early researcher who helped us understand much about charge , so we will give him a break!).

Types of Charge

We now have reason to believe that there are at least two types of charge, one for rubber and one for glass. But are there more?

No-rubbing demo

Let's start by introducing a new object, only this time we won't rub it with anything.

Now this is strange. The new item is attracted to both rods! What is going on? Have we discovered a new type of charge, one that attracts the other two types we have found?

Question 223.19.4

Maybe, but maybe the explanation of this phenomena is a little different. To understand this, let's consider how charge moves around.

Question 223.19.5

Movement of Charge

One of the strange things about charge is that it is *quantized*. We learned this word in when we found that only certain standing waves could be formed between boundaries. We are using this word in a similar way now. It means that charge has a smallest unit, and that it only comes in whole number multiples of that unit. Charge comes in a basic amount that can't be divided into smaller amounts. So like our standing wave frequencies, only certain amounts are possible As far as we know, the smallest amount

of charge possible is the electron charge.¹⁵ This charge we will call negative. We say that the electron is the principle charge carrier for negative charge. This fundamental unit of charge was found to be about

$$e = 1.60219 \times 10^{-19} \text{ C} \quad (19.5)$$

where the C stands for *Coulomb*, the *SI* unit of charge.

Any larger charge must be a multiple of this fundamental charge

$$Q = n \times e \quad (19.6)$$

The proton is the principle charge carrier for positive charge. From chemistry, you know protons are located in the nucleus of an atom, along with the neutron. In the Bohr model of the atom, the nucleus is surrounded by a cloud of electrons. The proton has the same amount charge as the electron (e), but is opposite in sign.

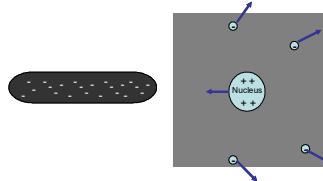
In a gram of mater, there are many, many, units of charge. There are about 5.0125×10^{22} carbon atoms in one gram of carbon. Each carbon atom has twelve protons and about twelve electrons. That is a lot of charge! But notice that the net charge is zero (or very close to it!). It is common for most mater to have zero net charge.

As far as we know, charge is always conserved. We can create charge, but only in plus or minus pairs, so the net charge does not change. We can destroy charge, but we end up destroying both a positive and a negative charge at the same time. The net charge in the universe does not seem to change much. So when something becomes charged, we expect to find that the charge has come from another object.

Lets go back to our rubber rod and glass rod demo. We rubbed the rod that was in our hand, but where did the charge come from? We believe that we are moving charge carriers (usually electrons) from one object to another, stripping them from their atoms. This happens when we use friction (rubbing) to charge the rods.

But what about our object that we did not rub, or our paper (we did not rub the bits of paper). We believe that charge can move, that is why scientists looked for and found charge carriers. Even in an atom, if I bring a charged object near the atom then the negative charge carriers (electrons) will experience a force directed away from the charged object, and the positively charged nucleus will experience a force pulling toward the charge object

¹⁵ I am not counting quarks here, which have a charge of $\frac{1}{3}$ or $\frac{2}{3}$ of the basic electron charge. But still, $\frac{1}{3}$ of the basic electron charge seems to be a real fundamental unit for quark based particles. And quarks aren't stable on their own, so we never see fractional charge in nature.



Notice that the electrons and the nucleus will *attract* each other, so the atom won't split apart. But it will become positively charged on one side because there are more positive charge carriers on that side. It will become more negatively charged on the other side, because there are more negative charge carriers on that side. We could draw the atom like this (figure 19.14) The force due to charge depends on how far away the charges are

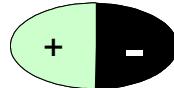
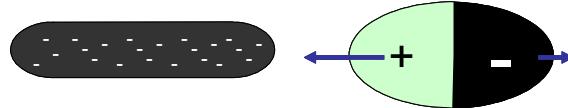
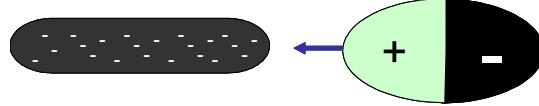


Figure 19.14.Polarized Atom

from each other. The attractive force between the positively charged side of the atom and the negative rod will have a stronger force than the negatively charged side of the atom and negatively charged rod will experience because the negative side is farther away. We will say that the atom has become *polarized*.

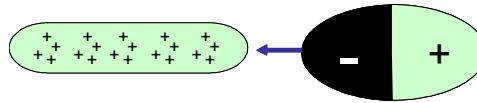


The positive side will experience an attractive force. The negative side will experience a repelling force. The net force due to the charge will be an attractive force. The atom will be accelerated toward the rod! We have seen something like this before. Remember an object in a fluid experiences a downward pressure force on the top, and an upward pressure force on the bottom. The pressure force is larger on the bottom, so there is an upward Buoyant force. The case with our polarized atom is very similar. We have a net electrical attractive force.



Now suppose we have lots of atoms (like our uncharged object or our bits of paper). Will they be attracted to the rod? Yes!

How about if we use a glass rod?



Everything is the same, only we switch the signs. The glass rod is positively charged. It will attract the electrons, and repel the nucleus. The atom becomes charged. The net force is attractive (positive rod and closer negative side of the atom)

We sometimes call the separation of charge in an insulator *polarization*.

Flow of Charge

Salt Shaker Demo

Let's start by introducing a new object, a salt shaker (my salt shaker is glass with a metal top). We will rub the salt shaker and see if it gets charged by placing it next to our charged rods.

Question 223.19.6

Now this is strange. We rubbed the object, but it was attracted to both rods as if there were no charge. We know glass can be charged. What is the problem?

Metal Demo

It turns out that some materials allow charge carriers to flow through them. Our experience with the lighting in our house might suggest that metals will do this. Let's try some other metal objects and see what we find.

It seems that the atoms are not maintaining a charge separation in these metal atoms! Some materials allow charge carriers to move through them. Usually these materials are metals, but most materials will allow some charge to go through them-even you-which is what is happening in this case. I charge the rod, but the charge leaves through my body. Other materials resist the flow of charge. Materials that allow charge to flow are called *conductors*. Materials that resist the flow of charge are called *insulators*.

Charging by Induction

Knowing that charge carriers can flow though a material, we can think of a way to charge a conductor. Lets suspend a conducting rod.



It is not initially "charged" meaning that it has the same number of positive charges and

negative charges, and they are evenly mixed together. I will bring a charged rod next to it.



but let's attach a wire to the other end of the rod to allow the charge to flow away from our conducting rod. We will connect the rod to the ground (in this case, to a water pipe) because the ground seems to be able to accept large amounts of charge carriers. So the charge carriers will flow to the ground.

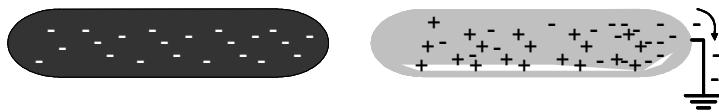
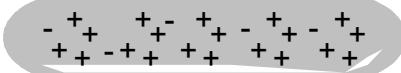


Figure 19.15.

(The strange little triangular striped thing is the electronics sign for a connection to the ground)

Now let's disconnect the wire from the rod. Is there a net charge on the conducting rod?



The answer is yes, because we now have more positive charges in the conducting rod than we have negative charges, so the net charge is positive.

Charging by Conduction

Suppose instead, I perform the same experiment, but I touch the rods. Now charge carriers can flow. Starting with an uncharged conductor,



I again bring in a charged rod. Again the charges separate in our conducting rod.

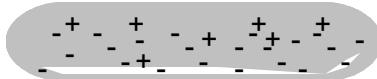


Then we touch the two rods. The excess charge on our charged rod flows to the conductor. Since in our drawing, the excess charge is negative, then some of the positive charge on the conductor is neutralized.



Take home lab assignment (using Scotch Brand Tape)

When we separate the rods, our conducting rod will have an excess of negative charge.



Notice that there is something different in our study of this new force. In the past, it was easy to tell which object was creating the environment and which was the mover. The Earth, being so much larger than normal objects, was the environmental object creating the gravitational acceleration that balls and cars and people move it. Then the balls and cars and people were the movers. Generally the thing causing the force, the environmental object, was much bigger than the mover. That is not true in our charge experiments so far. The rods are about the same size. So which is the environmental object and which is the mover? We will have to pick one to be our environmental object, and the other to be our mover. Sometimes the context of the problem helps. If the problem you are solving asks for the motion or the force on the rod on the right side of the diagram, then it is the mover and the rod on the left is the environmental object. If one charge is much larger than the other, we might be justified in calling this large charge the environmental object and a smaller charge near the big charge would be the mover.

Basic Equations

The minimum angle between two objects that can be resolved (according to the Rayleigh criteria) is

$$\theta_{\min} = 1.22 \frac{\lambda}{D}$$

20 Electric charge

Fundamental Concepts

- We have a model for how charge acts. The model tells us there are two types of charge, and that charges of similar type repel and charges of different type attract.
- We call the types of charge “positive” and “negative”
- In metals, the valence electrons are free to move around. We call materials where the charges move “conductors.”
- Materials where the valence electrons cannot move are called “insulators.”
- In insulators, the atoms can “polarize.”

Charge

Question 223.20.1

Question 223.20.2

Let's summarize what we tried to learn last time:

Question 223.20.3

Question 223.20.4

Model for Charge	
Frictional forces can add or remove charge from an object	
There are two, and only two kinds of charge	
Two objects with the same kind of charge repel each other	
To objects with different kinds of charge attract each other	
The force between two charged objects is long ranged	
The force between two charged objects decreases with distance	
Uncharged objects have an equal mix of both kinds of charge	
There are two types of materials, conductors (in which charges can move) and insulators (in which charges are fixed in place)	
Charge can be transferred from one object to another by contact between the two objects	

A serious shortcoming of this model is that it does not tell us what charge is. This is a shortcoming we will have to live with. We don't know what charge is any more than we can say exactly what mass or energy are. Charge is fundamental, as far as we can tell. We can't find a way to change charge into something else or to change something else into charge. For fundamental particles (like protons and electrons) either a particle has

charge, or it does not.

Conservation of charge

In some ways, this is really great! We have a new quantity that does not ever change. We can say that charge is conserved in the universe. Like energy, we can move charge around, but we don't create or destroy it. When we rubbed the plastic rods with rabbit fur or wool, we were removing charge that was already there in the atoms of the fur. If you take PH279 you might find that there are some caveats to this rule. We can make positron and electron pairs from high energy gamma rays. But when we do this we must always make a pair; one positive, and one negative. So the net charge remains unaffected.

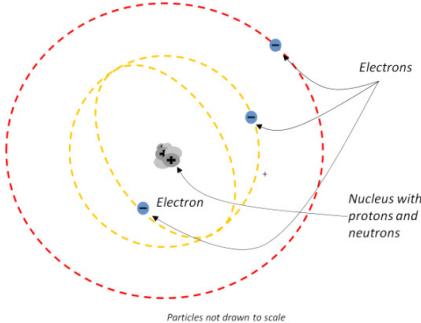
Insulators and Conductors

Question 223.20.5

Let's return to charges and atoms. We have an intuitive feeling for what is a conductor and what is an insulator, but let's see why conductors act the way they do.

Potential Diagrams for Molecules

Back in high school or in a collage chemistry class you learned that electrons move around an atom.



In the figure there are two energy states represented. You may even remember the names of these energy states. The orange-yellow lines show one “orbital distance” for the electrons near the nucleus. The red line shows another electron at a larger orbital distance. The inner orbital is a $1s$ state and the outer orbital is a $2s$ state. If these were satellites orbiting the earth, you would recognize that the two orbits have different amounts of potential energy. This is also true for electrons in orbitals. If we plot the

potential energy for each state we get something that looks like this

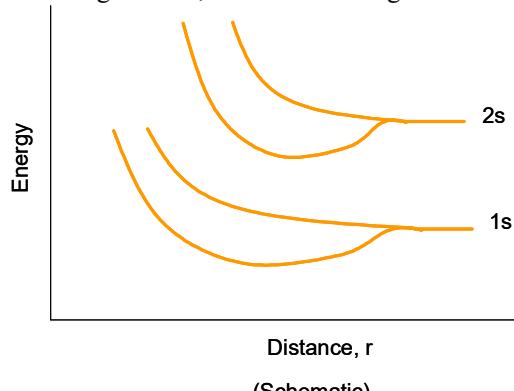


You can think of this as potential energy “shelves” where we can put electrons. If you were a advanced high school student, you learned that on the first two shelves you can only fit two electrons each. The higher shelves can take six, and so forth. But that won’t concern us in this class.

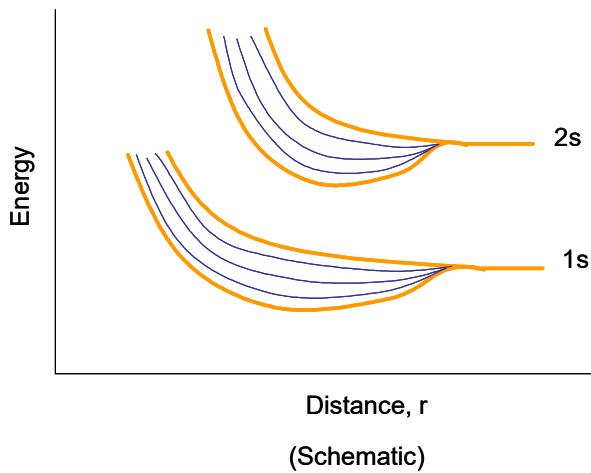
Building a solid

So far I have really only talked about single atoms. What happens when we bind atoms together? Let’s take two identical atoms. When they are far apart, they act as independent systems. But when they get closer, they start acting like one quantum mechanical system. What does that mean for the electrons in the atoms?

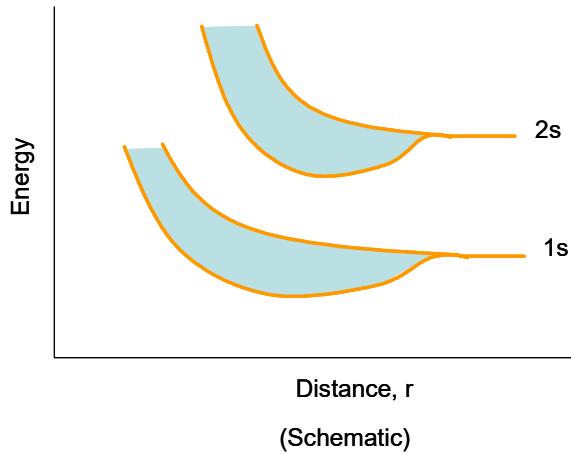
Electrons are funny things. They won’t occupy the exactly the same energy state. I can only have two electrons in a $1s$ state, but as I bring two atoms near each other I will have four! How does the compound solve this problem? The energy “shelves” split into more shelves. As the atoms get closer, we see something like this



At some distance, r , the states split. So each electron is now in a different state. Suppose we bring 5 atoms together.

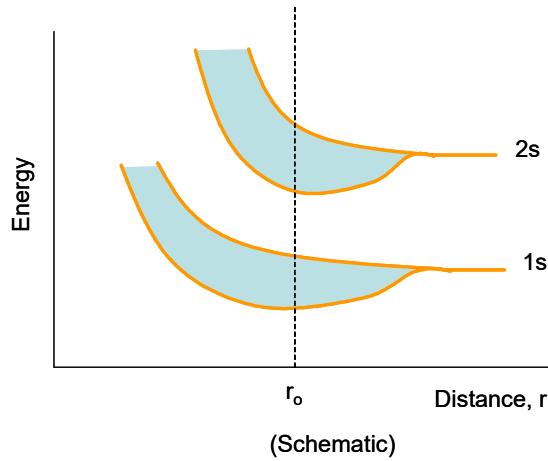


I get additional splitting of states. Now I have five different $1s$ states, enough for 5 atoms worth of $1s$ electrons. But solids have more than five atoms. Let's bring many atoms together.



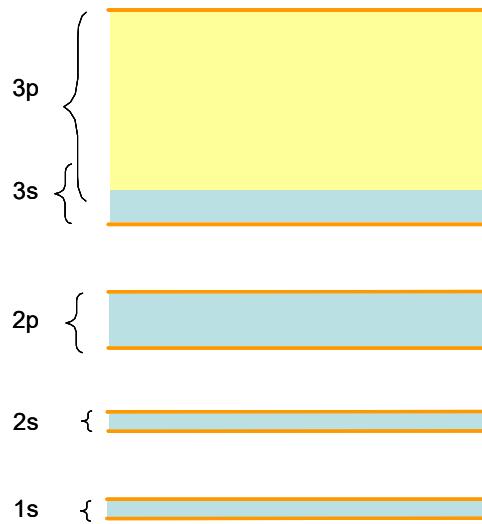
Now there are so many states that we just have a blue blur in between the original two split states. We have created a nearly continuous set of states in two bands. Each electron has a different energy, but those energy differences might be tiny fractions of a Joule. The former two states have almost become continuous bands of allowed energy states.

The atoms won't allow themselves to be too close. They will reach an equilibrium distance, r_o where they will want to stay.



(Schematic)

Since this is where the atoms usually are. We will not draw the whole diagram anymore. We will instead just draw bands at r_o (along the dotted line). Here is an example.



This means we have *bands* of energies that are allowed, that electrons can use, and *gaps* of energy where no electron can exist.

Conduction in solids

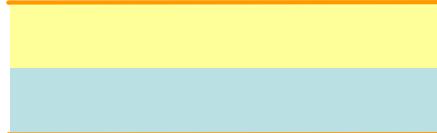
Notice that in our last picture, the 3s and 3p bands have grown so much that they overlap. The situation with solids is complicated. Notice also that the lower states are

blue. We will let blue mean that they are filled with electrons taking up every available energy state. The upper states are only partially filled. Yellow will mean the energy states are empty. We will call the highest completely filled band the *valence band* and the next higher empty band the *conduction band*.

We have three different conditions possible.

Metals

In a metal, the highest occupied band is only partially filled

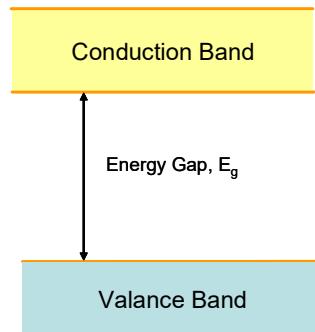


the electrons in this band require only very little energy to jump to the next state up since they are in the same band and the allowed energies are very closely spaced. Remember that movement requires energy. So if I connect a battery to provide energy, the electrons must be allowed to gain the extra energy, kinetic energy in this case, or they will not move. But in the case of a metal, there are easily accessible energy states, and the electrons flow through the metal.

We can say that the outer electrons are shared by all the atoms of the entire metal, so the electrons are easy to move for metals.

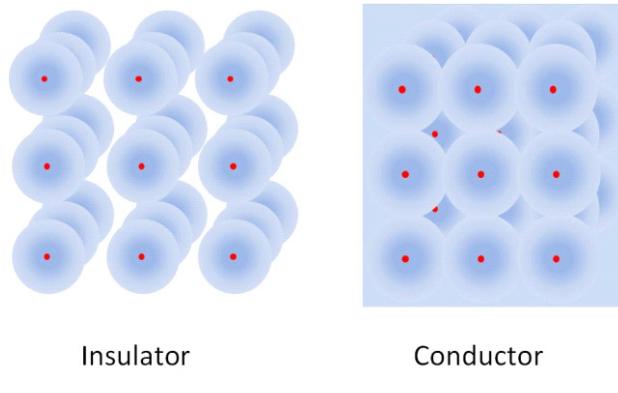
Insulators

A second condition is to have a full valence band and an empty conduction band. The bands are separated by an energy gap of energy E_g .



In this case, it would take a whopping big battery to make the electrons move. The battery would have to supply all of the gap energy plus a little more to get the electron to move. You might envision this as if there were an electrical “glue” that keeps the electrons in place. Before they can move, you have to free them from the “glue.” It takes an amount of energy, E_g , to free the electrons before they are able to accept kinetic energy. If we do connect a very large battery, say, 33000 V, then we can get electrons to jump the gap to a higher energy “shelf.” But high voltages are not normal conditions, so this is not usually the case. A material that has a large energy gap between its valence band and an empty conduction band is called an insulator.

A mental picture for this might be as shown in the next figure.

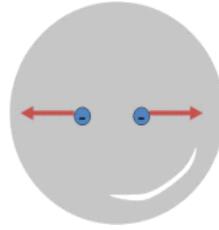


Question 223.20.6

Question 223.20.7

The insulator atoms keep their valence electrons bound to the nuclei of the atoms. But for a conductor, the valence electrons are free to travel from atom to atom.

In an isolated conductor, normally the charge is balanced, so the electrons may move but generally they stay near a nucleus. But if a conductor has extra electrons, the electrons that can move will move because they repel each other. So any extra charge will be on the surface of the conductor.



This happens very quickly, generally we do find the extra charge distributed on the outside of a conductor.

Semiconductors

The third choice is that there is a band gap, but the band gap is small. In this case, some electrons will gain enough thermal energy to cross the gap. Then these electrons will be in the conduction band. Devices that work this way are called semiconductors. We won't deal with semiconductors much in this class, but you probably used many of them in ME210. Diodes, and transistors are made from semiconductors.

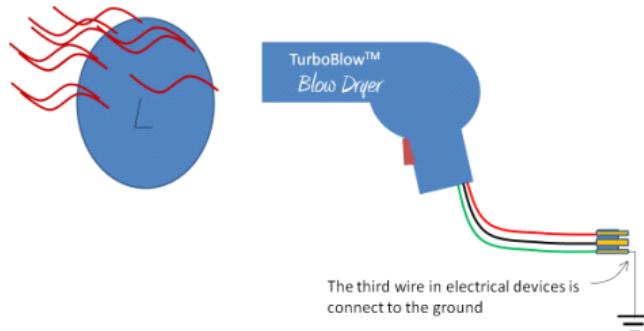
Charging and discharging conductors

Conductors can't usually be charged by rubbing. The electrons in the conductor may move when rubbed, but then they are free to move around in the conductor, so they don't leave. But if we rub an insulator, the electrons are not free to travel in the insulator material, so we can break them free. Once this happens, we can take our charged insulator and place it in contact with a conductor. The charge can flow from the insulator to the conductor (and arrange itself on the conductor surface). Once the charge has moved to the exterior, it will reach what we call *electrostatic equilibrium*. All of the repelling electrical forces are in balance, so the charges come to rest with respect to the conductor.

Question 223.20.8

We can remove the extra charge by creating a path for the charge to follow. Consider charging a balloon by rubbing it on your hair. Then you connect a wire to the balloon that is also connected to a metal water pipe. The charge can flow through the metal conducting wire. If there is a large body that can attract extra charge, the charge will flow. The Earth is such a large body that can attract the extra charge. The charge will

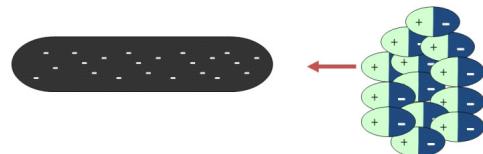
flow through the wire and pipe and go into the ground.



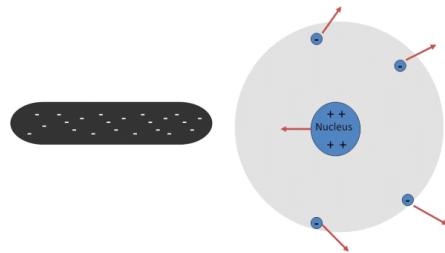
You may have heard of electrical grounds. This literally means tying your device to the Earth through a wire. Since you are made mostly of water that contains positive ions, you are also a conductor. So if we touch a charged object, we will most likely discharge the object. This is also why we must be careful with charge. Large amounts of charge flowing through us leads to death or injury.

If an object is *grounded*, it cannot build up extra charge. This is good for appliances and houses, and people.

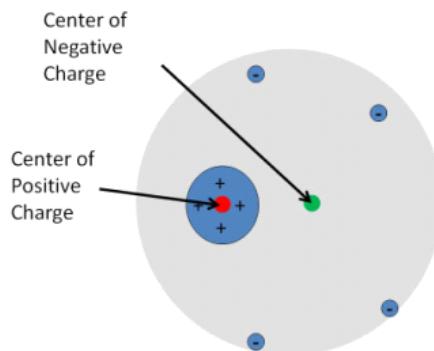
We talked last time about insulator atoms being polarized.



Remember that for each atom the electrons are displaced relative to the nucleus.

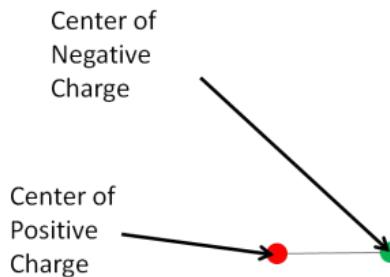


We can define a *center of charge* much like we defined a center of mass. In the case in the figure, we can define a negative center of charge and a positive center of charge.

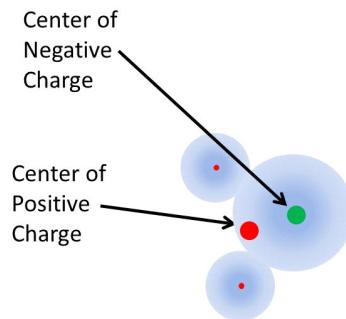


Question 223.20.9

Notice that the negative and positive center of charge are not in the same place when the atom is polarized. We have a name for a pair of positive and negative charges that are separated by a distance, but that are still bound together. We call it an *electric dipole*. Often we just draw the centers of charge joined by a line.



Using this we can explain why humidity affects our last lecture experiments so much. The water molecule has two hydrogen atoms and one oxygen atom. The covalent bond between the oxygen and hydrogen atoms forms when the oxygen “shares” the hydrogen’s electrons. The electrons from the hydrogen atoms spend their time with the oxygen atom making one side of the molecule more positive and the other side more negative.



Thus if you have a charged balloon on a humid day, one side of the water molecules in the air will be attracted to the extra charge on the balloon. The extra charge will attach

to the water molecules, and float away with them. This will discharge the balloon.

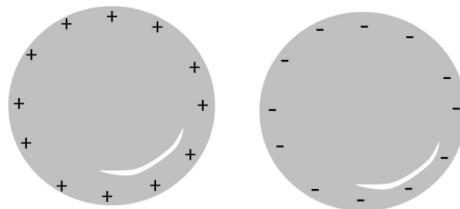
Note on drawing charge diagrams

We will have to draw diagrams in our problem solutions. Normally we won't draw atoms, so we will be drawing large objects with or without extra charge. We know that all materials have positive nuclei and negative electrons. When these are balanced, there is an electron for every proton, so if we add up the charges we get zero net charge. These charges don't contribute to net forces because for every attraction there is a repulsion of equal magnitude.

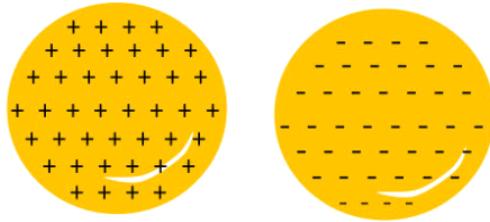
So we won't draw all of these charges, but we should remember they are there. We usually draw a cross section, so here is the cross section of a round, conducting ball.



But if we have extra charge, we should draw it. We will just add plus signs or minus signs. We won't draw little circles to show the electrons (we can't draw them to scale, they are phenomenally small). Here is an example of two round objects, one positive and one negative



If the objects are not conductors, the extra charge may be spread out. We draw the charge throughout the cross section of the object.



Note that if you transfer charge, from one object to another, you should try to keep the same total number of “+” or “-” signs to show the charge is conserved.

Basic Equations

21 Coulomb's Law and Lines of Force

Fundamental Concepts

- Our “charge” force is called the Coulomb force, and is given by $F = k_e \frac{|q_1||q_2|}{r^2}$
- A field is a quantity that has a value (magnitude and direction) at every point in space
- The Coulomb force is caused by an electric field
- We use field lines to give ourselves a mental picture of a field

Coulomb's Law

My experience so far is that Statics and Dynamics did not teach Newton's law of gravitation so teach it here.

Question
223.21.0.1

Question
223.21.0.2

Question
223.21.0.3

Question
223.21.0.4

Sometime ago in your Dynamics or PH121 class you learned about gravity. Let's review for a moment.

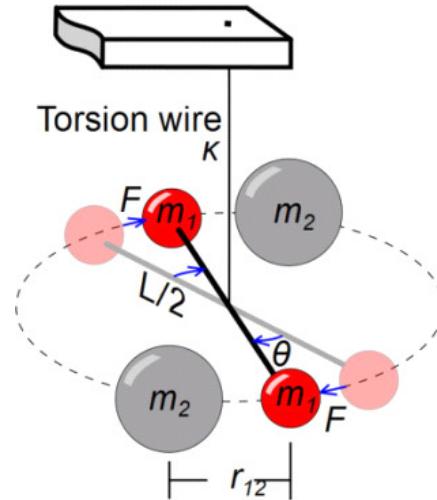
From our experience we know that more massive things exert a stronger gravitational pull than less massive things. We also have some idea that the farther away an object is, the less the gravitational pull. Newton expressed this as

$$F_g = G \frac{m_1 m_2}{r_{12}^2}$$

where the two masses involved (say, the Earth and you) are m_1 and m_2 and the distance between the two masses is r_{12} (e.g. the distance from the center of the Earth to the center of you). The constant G is a constant that puts the force into nice units that are convenient for us to use, like newtons (N) . It has a value of

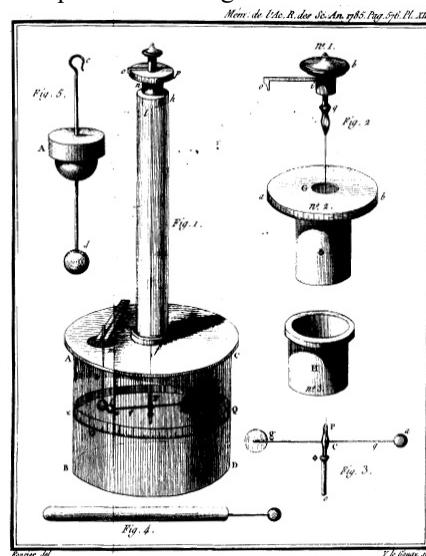
$$G = 6.67428 \times 10^{-11} \frac{\text{N m}^2}{\text{kg}^2}$$

You might ask, how do we know this? The answer is that Newton and others performed experiments. Newton's law of gravitation is empirical, meaning that it came from experiment. Lord Cavendish used a clever device to verify this law. He suspended two masses from a wire. Then he placed two other masses near the suspended masses.



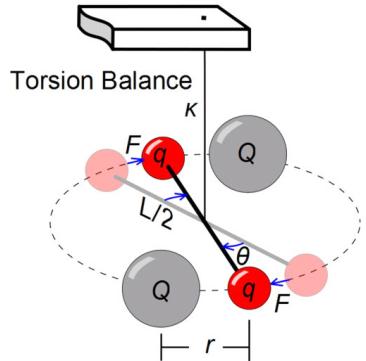
He knew the torsion constant of the wire (how much it resists being twisted). Then by observing how far the suspended masses moved, he could work out the strength of the gravitational force. This is called a torsion balance.

Charles Coulomb thought he could use the same device to measure the strength of the electric force. Here is his experimental design.



Coulomb's Torsion Balance Apparatus

You can see this is really just a torsion balance. This time objects with equal mass *and equal charge* are suspended on either end of a rod. The rod is hung on a wire. Two other charges are brought an equal distance, r_{12} , from the other charges. Knowing the torsional properties of the wire, the force due to the charges can be found.



Coulomb determined that the force due to a pair of charges has the following properties:

1. It is directed along a line connecting the two charged particles and is inversely proportional to the distance between their centers
2. It is proportional to the product of the magnitudes of the charges $|q_1|$ and $|q_2|$.
3. It is attractive (the charges accelerate towards each other) if the charges have different signs, and is repulsive (the charges accelerate away from each other) if the charges have the same signs.

We can write this in an equation

$$F = k_e \frac{|q_1| |q_2|}{r_{12}^2} \quad (21.1)$$

Question 223.21.2

Note how much this looks like gravitation! In the denominator, we have the distance, r_{12} , between the two charged particles' centers. We have two things in the numerator. But now we have $|q_1|$ and $|q_2|$ instead of m_1 and m_2 . We have a constant k_e instead of G , but the equation is very much like Newton's law of gravitation. That should be comforting, because we know how to use Newton's law of gravitation from PH121 or Dynamics. There is a very big difference, though. Gravitation can only attract masses, The Force due to charges can attract *or repel*.

Again there is a constant to fix up the units. Our constant is

$$k_e = 8.9875 \times 10^9 \frac{\text{N m}^2}{\text{C}^2} \quad (21.2)$$

which allows us to use more meaningful units (to us humans) in the force equation.

Comb and paper bits demo

How about strength? Is gravity or is this force due to charge stronger?

Force	Varies with Distance	Attracts	Repels	Acts without contact	Strength
Gravity	Yes	Always	Never	Yes	Weaker
Charge Force	Yes	Sometimes	Sometimes	Yes	Stronger

Lets try an example problem:

Example 21.1 Calculate the magnitude of the electric force between the proton and electron in a hydrogen atom. Compare to their gravitational attraction. We expect the electrical force to be larger. We need some facts about Hydrogen

Item	Value
Proton Mass	$1.67 \times 10^{-27} \text{ kg}$
Electron Mass	$9.11 \times 10^{-31} \text{ kg}$
Proton Charge	$1.6 \times 10^{-19} \text{ C}$
Electron Charge	$-1.6 \times 10^{-19} \text{ C}$
Proton-electron average separation	$5.3 \times 10^{-11} \text{ m}$

then,

$$\begin{aligned} F_e &= k_e \frac{|q_1| |q_2|}{r^2} \\ &= 8.9875 \times 10^9 \frac{\text{N m}^2}{\text{C}^2} \frac{(-1.6 \times 10^{-19} \text{ C})(1.6 \times 10^{-19} \text{ C})}{(5.3 \times 10^{-11} \text{ m})^2} \\ &= -8.1908 \times 10^{-8} \frac{\text{m}}{\text{s}^2} \text{ kg} \end{aligned}$$

and

$$\begin{aligned} F_g &= G \frac{m_1 m_2}{r^2} \\ &= 6.67 \times 10^{-11} \frac{\text{N m}^2}{\text{kg}^2} \frac{(1.67 \times 10^{-27} \text{ kg})(9.11 \times 10^{-31} \text{ kg})}{(5.3 \times 10^{-11} \text{ m})^2} \\ &= 3.6125 \times 10^{-47} \frac{\text{m}}{\text{s}^2} \text{ kg} \end{aligned}$$

which shows us what we expected, the gravitational force is very small compared to the electric force.

Permittivity of free space

It is customary to define an additional constant

$$\epsilon_o = \frac{1}{4\pi k_e} = 8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2} \quad (21.3)$$

Using this constant

$$F = \frac{1}{4\pi\epsilon_o} \frac{|q_1| |q_2|}{r^2} \quad (21.4)$$

which really does not seem to be an improvement. But if you go on to take an advanced class in electrodynamics you will find that this form is more convenient in other unit systems. So we will adopt it even though it is an inconvenience now.

Question 223.21.3

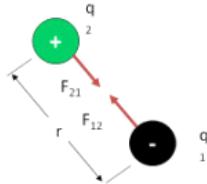
Question 223.21.4

Question 223.21.5

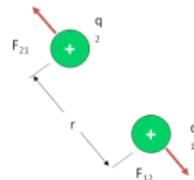
Direction of the force

What about direction? So far we have only calculated the magnitude of the force. But a force is a vector, so it must have a direction. Notice that our equation has absolute value signs in it. We will only get positive values from Coulomb's law.

To find a strategy for getting the direction, let's observe two charged objects



Experiments show that they seem to be pulled straight toward each other. The force seems to be along the line that passes through the center of charge for each of the two charged objects. We have to find this line from the geometry of our situation and our choice of coordinate systems. To make matters worse, we could have two of the same kind of charge.



The force will still be on the line connecting the centers of charge, but it will be in the opposite direction compared to the last case where the charges were of different sign. This seems complicated, and it is. We must observe the geometry of our situation and note whether the charges are the same or different signs to find the direction. Our equations can't tell us the direction on their own. You can't put the signs of the charges into the formula and expect a direction to come out! You have to draw the picture. Here is the process:

1. Define your coordinate system.
2. Find the line that connects the centers of charge. The force direction will be on that line.
3. Determine the direction by observing the signs of the charges. If the charges have the same sign, the force will be repulsive, if the charges have different signs, it will be attractive.

More than two charges

It is great that we know the force between two charges, but we have learned that there are billions of charges in everything we see or touch. It would be nice to be able to use our simple law of force on more than one or two charges. We did this with gravity.

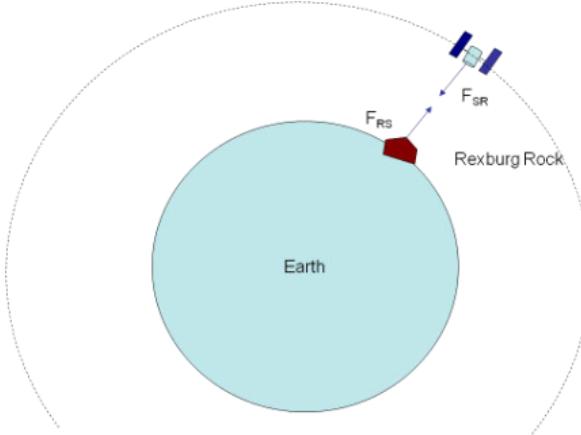
Question 223.21.6

Let's review.

Suppose I have a satellite orbiting the Earth. That satellite feels a force given by

$$\begin{aligned} F_g &= G \frac{M_E m_s}{r^2} \\ &= G \left(\frac{M_E}{r^2} \right) m_s \end{aligned}$$

but consider that on the Earth below the satellite, there is a rock on the surface of the Earth.



Part of the force due to gravity on the satellite must be due to this rock. We could write our force due to gravity as

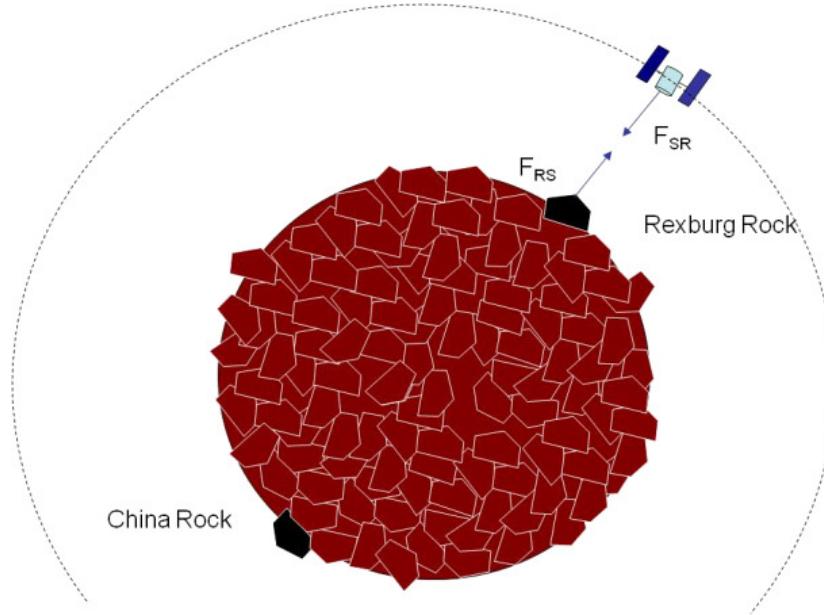
$$F_g = G \left(\frac{M_{rest}}{r_{rest}^2} \hat{r}_{rest} + \frac{M_{rock}}{r_s^2} \hat{r}_{rock} \right) m_s$$

where M_{rest} is the mass of all the rest of the Earth, minus the rock. If we take the Earth rock by rock, we would have

$$F_g = G \left(\sum_i \frac{M_i}{r_i^2} \hat{r}_i \right) m_s$$

where M_i is the mass of the i^{th} piece of the Earth and \hat{r}_i is the direction from M_i to m_s . We would not really want to do this calculation, because it would take a long time. Instead, back in PH121 or Dynamics we found we could add up all the mass and treat the Earth as one big ball of mass and represent it as if the mass was all at its center of mass (as long as there is no rotation so no torque). But let's think about all this mass. Does the force between a rock in China and our satellite get diminished because our

rock in Rexburg is in the way?



No, the force due to gravity is really the sum of all the little forces between all the parts of the Earth and our satellite. One bit of mass does not interfere with the force from another bit of mass.

Now let's look at the electric force. Suppose we have many charges in some configuration (maybe a round ball of charge). We could call the total charge, Q_E . Then our force magnitude on a mover charge q_m , would be

$$F_e = k_e \frac{|Q_E| |q_m|}{r^2}$$

The collection of charge Q_E would be the environmental charge. But we can picture this as the individual parts of Q_E all with little forces pairs acting on q_m summing up to get F_e .

$$F_e = k_e \sum_i \left(\frac{|Q_i|}{r_i^2} \hat{\mathbf{r}}_i \right) |q_m|$$

where Q_i is a piece of the total charge Q_E .

This is an amazingly simple idea. The force on a mover charge, q_m , due to any number of charges is just the sum of the forces due to each charge acting on q_m . Sometimes the mover charge is called a *test charge*, but we will call it a mover charge and we will call the Q_i environmental charges.

Suppose in our ball of charge, we have an element of charge on the opposite side of the ball and another element of charge close to us. Would the near charge element "screen

off" or some how reduce the force due to the far charge element?

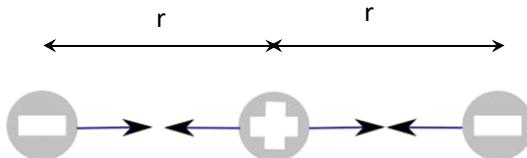
Like with gravity, it would not. Note that because one charge is farther away, the force from the far charge is not the same magnitude as that of the near charge. But we calculate both using our formula, and add them up (a vector sum) with all the others.

While we are talking about it, it might seem that the rest of the matter in the ball will screen off the electric force. But matter, itself, does not interfere with our electric force. Only other charges will change the force, and then only following the idea of that their forces add as vectors (remember that for electricity they can cancel, because we have both positive and negative charges).

Recall that in our study of waves, when we had two waves in a medium we found we could just added up displacements point for point. We called this *superposition*. We will use the same word here, but it has a slightly different meaning. We are not adding up wave displacements. We are adding up forces. But we still do it point for point.

Now where there are forces, there will be Newton's second law! Let's consider a problem. Suppose we have three charges, equally spaced apart as shown where each has the charge of one electron (q_e) but the middle charge is positive and the other two are negative

Draw picture on board



We identify the middle charge as the mover (since we are asked for the force on this charge) and the left and right charges as the environmental charges. We can draw a free body diagram for the mover charge.



and find the net force on the mover charge, then

$$\vec{F}_{net} = m \vec{a} = \vec{F}_R + \vec{F}_L$$

We only have x -components so we can write this as

$$F_{net_x} = ma = F_{Rx} - F_{Lx}$$

where the minus sign is used for F_{Lx} because it is pointing to the left and that is usually the minus x direction.

We may ask, is this mover charge accelerating? We may suspect that the answer is no, but here we have something new. We don't know the magnitude of F_R or F_L . We now have to find the magnitudes to know. Back in PH121 you would have been given the magnitude of the forces, but in a charge problem we know how to calculate the magnitudes, so let's do that. We can use the formula for the Coulomb force

$$F = k_e \frac{|q_1||q_2|}{r^2}$$

we can use r as the distance from the middle charge to each of the other charges since in this special case they are both the same distance from the middle charge. Then

$$F_R = k_e \frac{q_e^2}{r^2}$$

$$F_L = k_e \frac{q_e^2}{r^2}$$

these are the magnitudes. We should notice that F_L points to the left. So we need to include a minus sign in front of its magnitude.

$$F_{net_x} = ma = F_{Rx} - F_{Lx}$$

$$F_{net_x} = ma = k_e \frac{q_e^2}{r^2} - k_e \frac{q_e^2}{r^2}$$

$$= 0$$

now we can say that the middle charge is definitely not accelerating.

Of course this is a pretty easy Newton's 2nd law problem. It was all in the x -direction. But suppose that is not true. Then we need to take components of the forces vectors.

Draw picture on board

Let's try one of those.



Here is a new configuration of our charges. There will be a Coulomb force between each negative charge the positive charge. What is the net force on the positive charge?

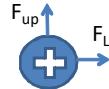
Again we need Newton's second law and the Coulomb force equation. We identify the positive charge as our mover, and the negative charges as the environmental charges. Our basic equations are

$$F = k_e \frac{|q_1||q_2|}{r^2}$$

Draw picture on board

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

but this time we need an x and a y Newton's second law equation. Let's draw the free body diagram. I have chosen the positive y -direction to be upward and the positive x -direction to be to the right.



The negative charge that is above our positive charge will cause an upward force. The negative charge to the right will cause a force that pulls to the right. This is a two-dimensional problem, so we need to split our Newton's second law into two one-dimensional problems.

$$F_{net_x} = ma_x = F_L$$

$$F_{net_y} = ma_y = F_{up}$$

so

$$F_{net_x} = k_e \frac{q_e^2}{r^2}$$

$$F_{net_y} = k_e \frac{q_e^2}{r^2}$$

We can see that there will be a force in both the x and the y direction. How do we combine these to get the net force? We use our basic equations for combining vectors:

$$\begin{aligned} F_{net} &= \sqrt{F_{net_x}^2 + F_{net_y}^2} \\ &= \sqrt{\left(k_e \frac{q_e^2}{r^2}\right)^2 + \left(k_e \frac{q_e^2}{r^2}\right)^2} \\ &= \sqrt{2} \frac{1}{r^2} k_e q_e^2 \end{aligned}$$

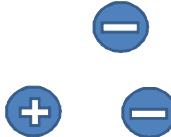
but we are not done. We need a direction. Generally we use the angle with respect to the positive x -axis.

$$\begin{aligned} \theta &= \tan^{-1} \left(\frac{F_{net_y}}{F_{net_x}} \right) \\ &= \tan^{-1} \left(\frac{k_e \frac{q_e^2}{r^2}}{k_e \frac{q_e^2}{r^2}} \right) \\ &= \frac{\pi}{4} \text{ rad} \end{aligned}$$

so we have a net force of $F = \sqrt{2} \frac{1}{r^2} k_e q_e^2$ at a 45° angle with respect to the x -axis.

Of course, this is still fairly simple, we should also review taking components of vectors that are not directed along the x and the y axis. Suppose we move the top charge as

Draw picture on board shown below



Draw picture on board

Once again the positive charge is the mover and the negative charges are the environment. Now our free body diagram looks like this:



Once again we have a two-dimensional problem. We need to convert it into two one-dimensional problems.

$$\begin{aligned} F_{net_x} &= ma_x = F_{L_x} + F_{2x} \\ F_{net_y} &= ma_y = F_{L_y} + F_{2y} \end{aligned}$$

but we don't know F_{L_x} , F_{2x} , F_{L_y} , and F_{2y} . But our basic equations should include how to make vector components

$$\begin{aligned} v_x &= v \cos \theta \\ v_y &= v \sin \theta \end{aligned}$$

where θ is measured from the positive x -axis. So

$$\begin{aligned} F_{net_x} &= ma_x = F_L \cos \theta_L + F_2 \cos \theta_2 \\ F_{net_y} &= ma_y = F_L \sin \theta_L + F_2 \sin \theta_2 \end{aligned}$$

and we realize that

$$\theta_L = 0$$

and that

$$\begin{aligned} \cos(0) &= 1 \\ \sin(0) &= 0 \end{aligned}$$

so

$$\begin{aligned} ma_x &= F_L + F_2 \cos \theta_2 \\ ma_y &= 0 + F_2 \sin \theta_2 \end{aligned}$$

This gives the x and y components of the net force on the positive charge. Using our Coulomb force for the magnitudes, we have

$$\begin{aligned} F_{net_x} &= k_e \frac{q_e^2}{r^2} + k_e \frac{q_e^2}{r^2} \cos \theta_2 \\ F_{net_y} &= k_e \frac{q_e^2}{r^2} \sin \theta_2 \end{aligned}$$

I will tell you $\theta = \frac{\pi}{4}$ rad (or 45°). So we can find

$$\begin{aligned} F_{net_x} &= k_e \frac{q_e^2}{r^2} + k_e \frac{q_e^2}{r^2} \left(\frac{\sqrt{2}}{2} \right) = k_e \frac{q_e^2}{r^2} \left(1 + \frac{\sqrt{2}}{2} \right) \\ F_{net_y} &= k_e \frac{q_e^2}{r^2} \left(\frac{\sqrt{2}}{2} \right) \end{aligned}$$

and

$$\begin{aligned} F_{net} &= \sqrt{F_{net_x}^2 + F_{net_y}^2} \\ &= \sqrt{\left(k_e \frac{q_e^2}{r^2} \left(1 + \frac{\sqrt{2}}{2} \right) \right)^2 + \left(k_e \frac{q_e^2}{r^2} \left(\frac{\sqrt{2}}{2} \right) \right)^2} \\ &= \frac{k_e q_e^2}{r^2} \sqrt{2 + \sqrt{2}} \end{aligned}$$

This is not so nice and easy. The angle is

$$\begin{aligned} \theta &= \tan^{-1} \left(\frac{k_e \frac{q_e^2}{r^2} \left(\frac{\sqrt{2}}{2} \right)}{k_e \frac{q_e^2}{r^2} \left(1 + \frac{\sqrt{2}}{2} \right)} \right) \\ &= \tan^{-1} \left(\frac{1}{2} \frac{\sqrt{2}}{\frac{1}{2}\sqrt{2} + 1} \right) \\ &= 0.39270 \text{ rad} \end{aligned}$$

Note that I am using symbols as long as I can. This will become ever more important in this course. The problems will become very complicated. It is easier to make mistakes if you input numbers early.

Also notice that I carefully placed the charges the same distance, r , from each other. Of course that will not always be true. If the distances are different, we will use subscripts (e.g. r_1, r_2) to distinguish the distances.

Fields

If you are taking PH223 you should have already taken PH121 or an equivalent class. In PH121, you learned about how things move. You learned about forces and how force relates to acceleration

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

The force, \vec{F} , is how hard you push or pull. This push or pull changes the motion of the object, represented by its mass, m . The change in motion is represented by its acceleration, \vec{a} . Notice that both \vec{F} and \vec{a} are vectors. We will need all that you

learned about vectors in PH121.

Since physics is the study of how things move, we are going to study the motion of objects again in this class. But in this section of our class we will learn about new sources of force, that is, new ways to push or pull something.

Really these new sources of force are not entirely new. You have heard of them and probably experienced them. They are electrical charge and magnetism. You have probably had a sock stick from you after pulling it out of a dryer, and you have probably had a magnet that sticks to your refrigerator. So although these new sources of force are new to our study of physics, they are somewhat familiar in every day lives.

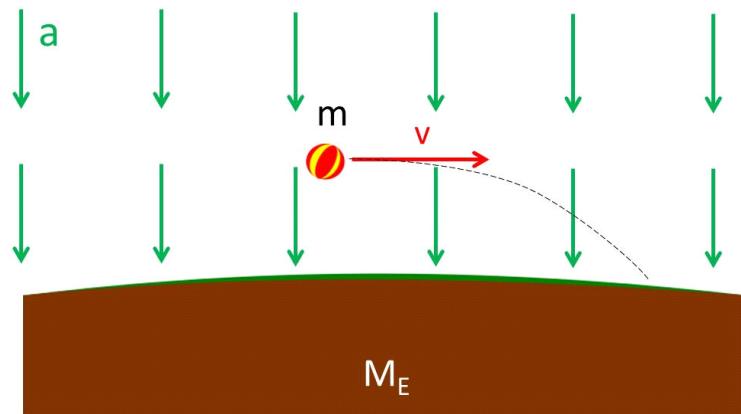
Let's review a particular force, the force due to gravity. This makes sense to do because our equation for electrical force is so very much like Newton's equation for gravity. Think of most of our experience with gravity. We have an object moving near the Earth. There is a force acting on the object, and that force is because of Earth's gravity.

We can think of the Earth as creating an environment in which the object moves, feeling the gravitational force. This is a property of all non-contact forces.

Think of a ball falling. We considered this as an environment of constant acceleration. In this environment, the ball feels a force proportional to its mass

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

where $g = 9.81 \frac{\text{m}}{\text{s}^2}$ is the acceleration due to gravity. This is true anywhere near the Earth's surface. We could draw this situation as follows:



where the environment for constant acceleration is drawn as a series of arrows in the

acceleration direction (downward toward the center of the Earth). Anywhere the ball goes the environment is the same. So we draw arrows all around the ball to show that the whole environment around the ball is the same.

Notice that the environment is described by an acceleration, g given by

$$\vec{g} = \frac{\vec{F}}{m}$$

that is, the environment is described by the force per unit mass.

This environment is caused by the Earth being there. If the Earth suddenly disappeared, then the acceleration would just as suddenly go to zero. So we can say that the Earth creates this constant acceleration environment.

Notice that there are two objects involved, the ball and the Earth. Also notice that one object creates an environment in which the other object moves. In our case, the Earth created the environment and the ball moved through the environment. This situation will recur many times, so let's give the objects these names, the Earth as the "Environmental object", and the ball as the "mover."

We should ask ourselves, does something like this happen with our electrical force. The electric forces is also a non-contact force. Could we view one charge as creating an environment in which the other charge moves? And if there is an environment, what would that environment be. Would it be an acceleration, or something else?

Michael Faraday came up with answers to this questions. To gain insight into his answers, let's consider our force again.

$$F_e = k_e \frac{|Q_E| |q_m|}{r^2}$$

but let's take q_m as a very small test charge that we can place near a larger distribution of charge Q_E . This is like the Earth and our small ball. The large Q_E is the environmental charge and the small q_m is the mover charge. We want q_m to be so small that it can't make any of the parts of Q_E rearrange themselves or any of the atoms forming the body that is charged with Q_E to polarize. Then we define a new quantity

$$\vec{E} = \frac{\vec{F}}{q_m}$$

This is the force per unit charge. This is very like our gravitational acceleration which is a force per unit mass. Then

$$\vec{F} = q_m \vec{E} \tag{21.5}$$

This is really like

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

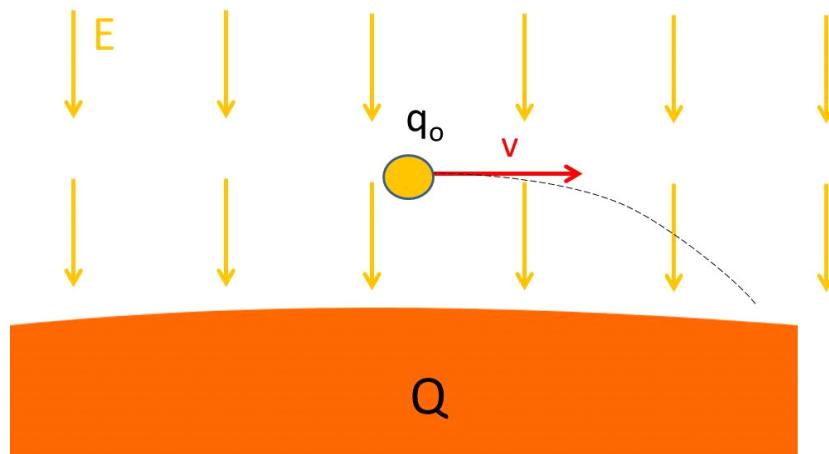
but with the mass replaced by q_m and the acceleration replaced by this new force-per-unit-charge thing. For gravity it is the mass that made the gravitational pull. With the electric force it is the charge that creates the pull. So replacing m with q_m makes some sense. But what does it mean that the acceleration has been replaced by \vec{E} . Well, since \vec{g} was the representation of the environment, can see that this new quantity is taking the place of the environment, but it can't be an acceleration. It does not have the right units. Let's investigate what it is.

Let's write the magnitude of E

$$\begin{aligned} E &= \frac{F}{q_m} \\ &= \frac{k_e \frac{|Q_E||q_m|}{r^2}}{q_m} \\ &= k_e \frac{Q_E}{r^2} \end{aligned}$$

van de Graff and test charge

But this is really not a quantity that we have seen before. It depends on how far away we are from the environmental charge Q_E . It has a value at every point in space—the whole universe! (think of our acceleration environment being all around the moving ball) though its values for large r are very small. The quantity is only large in the near vicinity of the charge, Q_E .



We can picture this quantity as being like a football field with something (an environmental charge) hidden out there on the grass. If we know where the object is, we can tell a searcher how “warm” or “cold” they are as they wander around looking for the object. For every location, there is a value of “warmness.” If we extend this idea to three dimensions, we are close to a picture of \vec{E} . The environment quantity \vec{E}

A field is a quantity that has a value (magnitude and/or direction) at every point in space.

Question 223.21.7

has a value at every point in three dimensional space. Since this is a new quantity, we need to give it a name. We will call it an *electric field*. But we have to add one more complication. It is a vector, so it also has a direction at each point in space as well. This direction is the direction the force would be on q_o , the mover, if we placed it at that location.

But where does this field come from? We say that an environmental charge Q_E creates a field

$$\vec{E} = k_e \frac{Q_E}{r^2} \hat{r} \quad (21.6)$$

centered at the charge location. The field is our environment for our mover.

Now we can understand our classical model about how gravity works! Have you wondered how a satellite knows that the Earth is there and that it should be pulled toward the Earth? We can say the Earth sets up a *gravitational field* because it has mass. The gravitational field shows up as an acceleration field. The satellite (the mover) feels the gravitational field because the field exists at the location of the satellite (it exists at all locations, so it exists at the satellite's location). The satellite does not have to know that the Earth is there, because it feels the field right where it is. The satellite reacts to the field, not directly with the Earth that created the field.¹⁶

Likewise, our charge Q_E has the property of creating an *electric field* as the environment around it. Other charges (movers) will feel the field at their locations, and therefore will feel a force due to the field created by Q_E .

Field Lines

Question 223.21.10

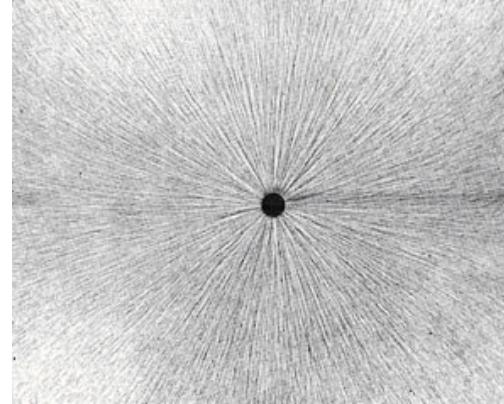
Magnet and Iron Filings

We need a way to draw the environment created by the environmental charge Q . We could draw lots of arrows like in the previous pictures, and we will do this sometimes. But there is another way to draw the environment that has become traditional. Have you ever taken iron filings and placed a magnet near them? If you do, you will notice that the filings seem to line up.

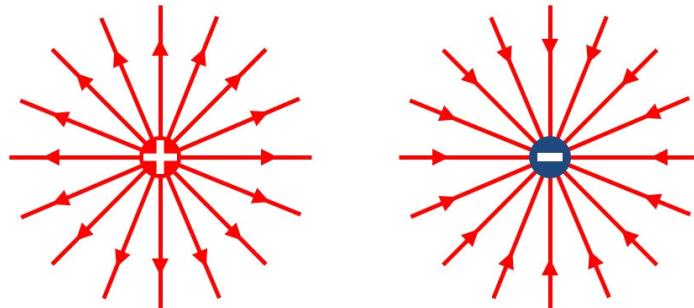
If you took PH121 you probably heard that there is a magnetic force. It is a non-contact force, so we expect it has a *magnetic field*. The iron filings are aligning because they are acted upon by the field. It is natural to represent this field as a series of lines like the ones formed by the iron filings. We will do this in a few lectures!

¹⁶ Here I am taking a quantum mechanical view of gravity. In General Relativity, the “field” is space that is warped by the mass of the Earth.

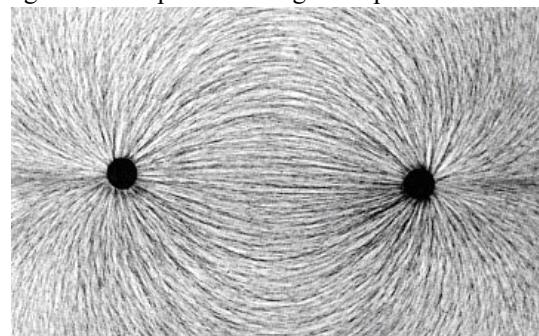
But there is a similar experiment we can do with the electric force. This is harder, but we can use small seeds or pieces of thread suspended in oil. These small things become polarized in an electric field. They line up like the iron filings.



http://stargazers.gsfc.nasa.gov/images/geospace_images/electricity/elec_field_lines.jpg
We can represent the electric field by tracing out these lines. The last figure would look like this

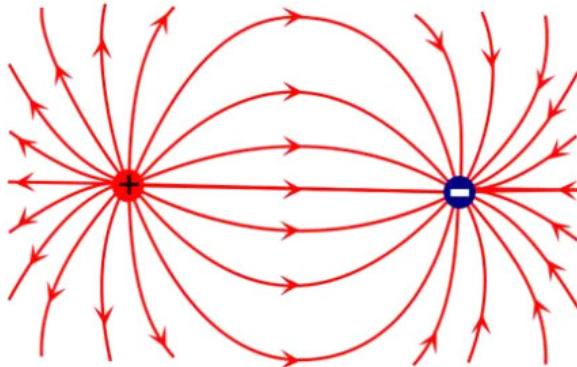


We can't tell if the charge was negative or positive from oil suspension picture, but if it was positive, by convention we draw the field lines as coming out of the charge. If it were negative the field lines would be drawn as going in to the charge. Here is a combination of a negative and a positive charge or dipole.

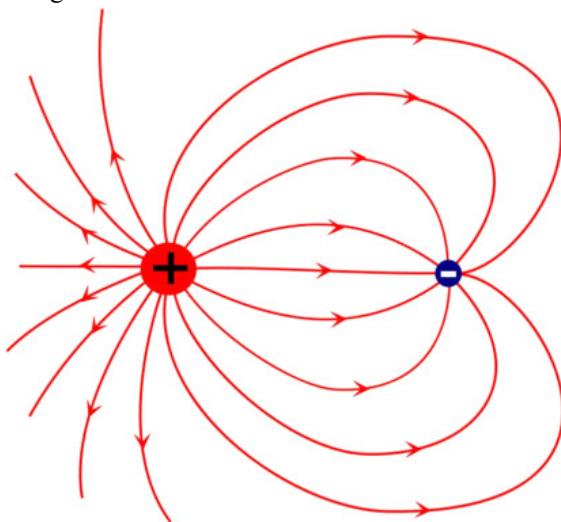


http://stargazers.gsfc.nasa.gov/images/geospace_images/electricity/elec_field_lines2.jpg

In this case both the positive and negative charges are working together to make the environment or field that a third charge could move through. The field line drawing would look like this.



This combination of positive and negative charges had equal charges, the only difference was the sign change. Here is one where the positive charge has more charge than the negative charge.



Notice that the number of field lines is proportional to the field, but there is no set proportionality. If the field from one charge is twice that of the other, we pick a number of field lines for, say, the negative charge, and double the lines on the larger positive charge.

This gives us a way to picture the electric field in our minds!

Some things to notice:

1. The lines begin on positive charges

2. The lines end on negative charges
3. If you don't have matching charges, the lines end infinitely far away (like the single charges in the first picture).
4. Larger charges have more lines coming from them
5. Field lines cannot cross each other
6. The lines are only imaginary, they are a way to form a mental picture of the field.

Question 223.21.11

We only draw the field lines for the environmental charges. Of course the mover charge also makes a field, but this self-field can't cause the mover charge to move. If it could we could have perpetual motion and that violates the second law of thermodynamics. Since the mover's self-field is not participating in making the motion, we won't take the time to draw it!¹⁷

Remember, field lines are not real, but are a nice way to draw the field made by the environmental charge. We will use field lines often in drawing pictures as part of our problem solving process.

On-Line demonstrations

An applet that demonstrates the electric field of point charges can be found here:

http://phet.colorado.edu/sims/charges-and-fields/charges-and-fields_en.html

If you prefer a video game, try Electric Field Hockey:

<http://phet.colorado.edu/en/simulation/electric-hockey>

As a wacky example of Coulomb forces, see this video of charged water droplets orbiting charged knitting needles on the Space Shuttle:

http://www.nasa.gov/multimedia/videogallery/index.html?media_id=131554451

Basic Equations

¹⁷ This picture will be a little more complicated when we allow for relativistic motion of charges and other more difficult effects, but that can wait for more advanced physics courses. For most engineering applications, this is a great approximation.

22 Electric Fields of Standard Charge Configurations Part I

Fundamental Concepts

- Adding of vector fields for point charges
- Standard configurations of charge

Standard Charge Configurations

Actual engineering projects or experimental designs require detailed calculations of fields using computers. These field simulations use powerful numerical techniques that are beyond this sophomore class. But we can gain some great insight by using some basic models of simple charged objects. We will often look at the following models:

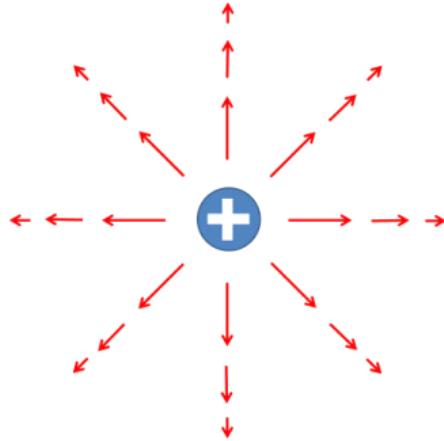
Standard Configurations of Charge
Point charge
Several point charges
Line of Charge
Semi-infinite sheet of charge
Charged sphere
Charged spherical shell
Ring of Charge

Point Charges

We have already met one of these standard configurations, the point charge

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r}$$

The field of the point charge is represented below



This picture requires a little explanation. The arrows are larger nearer the charge to show that the field is stronger. But note that each arrow is the magnitude and direction of the charge at one point. We really need a three dimensional picture to describe this, and even then the fact that the arrows have length can be misleading. The long arrows cover up other points, that should also have arrows. We can only draw the field at a few points, and at those points the field has both magnitude and direction. But we must remember that there is really a field magnitude and direction at every point.

To go beyond single charges we need a group of point charges of some sort. The fields add like forces

$$\vec{E} = \sum_i \vec{E}_i$$

where we recognize that we are summing vectors. Let's take a look at a few combinations of charges and find their fields

Two charges

Let's go back to our idea of an environmental charge, Q_E , and a mover charge, q_m . The mover charge is considered to be small enough that its effect on Q_E is negligible. So the field due to the large charge is unaffected by this small charge.

Of course, the total field is a superposition of both fields. We call the field produced by the little mover charge its *self-field*. But the mover charge can't move itself.¹⁸ The mover's self-field can't move the mover. So we don't draw the field due to q_m . We can envision an environmental field that is just due to the environmental charge, Q_E , as

¹⁸ This would allow perpetual motion, breaking the second law of thermodynamics.

Question 223.22.1

if there are no other charges anywhere in the whole universe. Of course this is not the case, but this is how we think of the field *due to* charge Q_E .

Question 223.22.2

We can identify that a charge q_m placed in this field due to Q_E will feel a force

$$\begin{aligned}\vec{F}_e &= q_m \vec{E} \\ &= \frac{1}{4\pi\epsilon_0} \frac{Q_E q_m}{r^2} \hat{r}\end{aligned}$$

due to the field

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{Q_E}{r^2} \hat{r}$$

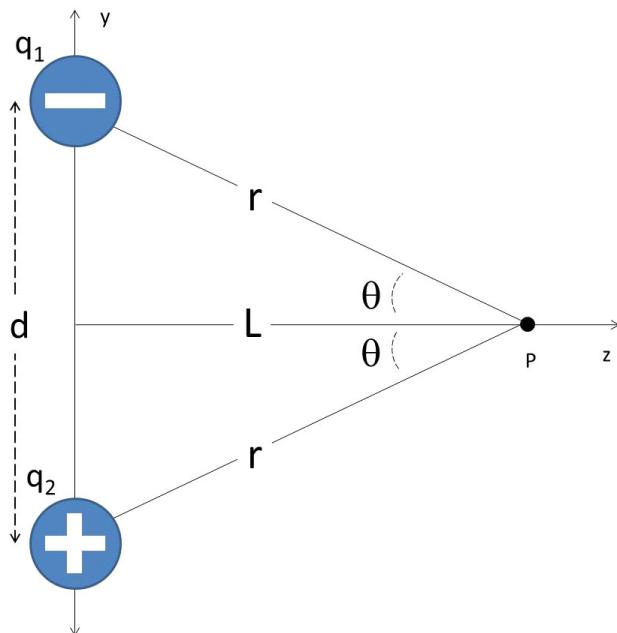
where this field is just due to Q_E and does not contain the contribution from q_m . So the charge q_m only feels a force due to the field created by charge Q_E . A third charge, q_{new} brought close to the other two would feel both \vec{E}_Q and \vec{E}_q . Then both Q_E and our original q_m would be environmental charges and the new charge q_{new} would be the mover. At this point, we would probably relabel Q_E and q_m as Q_1 and Q_2 and relabel q_{new} as q_m so we could tell that the original two charges are now the environment and the new charge is the mover.

Vector nature of the field

Question 223.22.3

Question 223.22.4

Remember that the field is a force per unit charge. Forces add as vectors, so we should expect fields to add as vectors too. Let's do a problem.

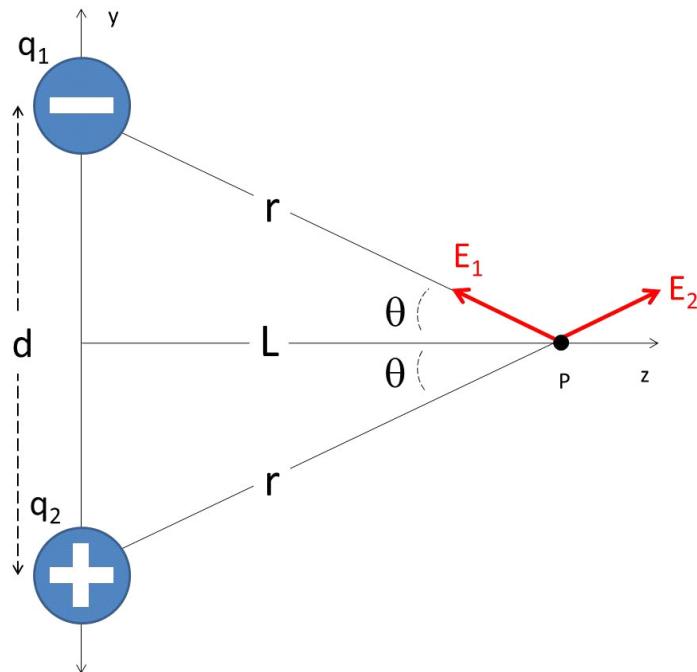


Two charges are separated by a distance d . What is the field a distance L from the center of the two charges?

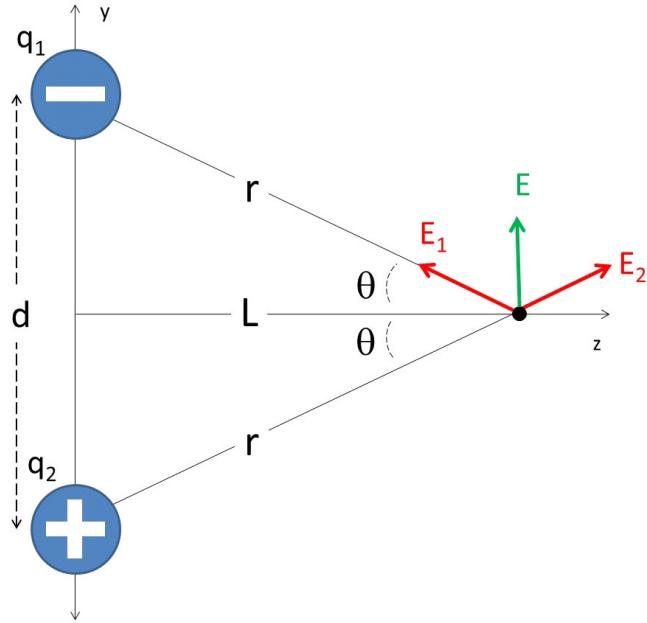
We should recognize this as our old friend, the dipole.

Note that both of these charges are environmental charges. We are asked in this problem to find the environment, the field. We don't really have a mover charge. But we could pretend we do have a mover, q_o at point P where we want to know the environment if it helps us picture the situation. But really we are calculating what the environment around the two charges will be.

We start by drawing the situation. I chose not to draw field lines. Instead I drew the field vectors at the point, P , where we want the field. The field lines would tell me about the whole environment everywhere, and that might be useful. But this problem only wants to know the field at one point, P . So it was less work to draw the field using vectors at our one point.



Note that I need a vector for each of the environmental charges. Each contributes to the environment. The contribution to the field due to environmental charge q_1 is labeled E_1 and likewise the contribution to the field from environmental charge q_2 is labeled E_2 .



The net environment is the superposition of the fields due to each of the environmental charges.

$$\vec{E}_{net} = \vec{E}_1 + \vec{E}_3$$

From the figure, we see that if we had a small mover charge, \$q_o\$ on the axis a distance at point \$p\$ then we would get two forces, one from each of the environmental charges \$q_1\$ and \$q_2\$. We can use Newton's second law to find the net force on our imaginary \$q_o\$.

$$F_{net_z} = ma_z = -F_1 \cos \theta + F_2 \cos \theta$$

$$F_{net_y} = ma_y = F_1 \sin \theta + F_2 \sin \theta$$

we can see that the distance from each charge to point \$P\$ is

$$r = \sqrt{\frac{d^2}{4} + L^2}$$

so

$$\sin \theta = \frac{d}{2\sqrt{\frac{d^2}{4} + L^2}}$$

we also know from Coulomb's law that

$$F_1 = F_2 = \frac{1}{4\pi\epsilon_0} \frac{qq_o}{r^2}$$

but we want the field, so we need to divide all of this by \$q_o\$

$$E_1 = E_2 = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$$

Our Newton's second law becomes an equation for the components of the combined

electric field.

$$\begin{aligned}\frac{F_{net_z}}{q_o} &= -\frac{F_1}{q_o} \cos \theta + \frac{F_2}{q_o} \cos \theta \\ \frac{F_{net_y}}{q_o} &= \frac{F_1}{q_o} \sin \theta + \frac{F_2}{q_o} \sin \theta\end{aligned}$$

or just

$$\begin{aligned}E_{net_z} &= -E_1 \cos \theta + E_2 \cos \theta \\ E_{net_y} &= E_1 \sin \theta + E_2 \sin \theta\end{aligned}$$

We can see from the figure that in the z -direction we will have no net field,

$$E_z = -E_1 \cos \theta + E_1 \cos \theta = 0$$

But in the y -direction we have

$$\begin{aligned}E_y &= E_1 \sin \theta + E_2 \sin \theta \\ &= 2E_1 \sin \theta \\ &= \frac{2}{4\pi\epsilon_o} \frac{q}{r^2} \sin \theta\end{aligned}$$

and since we found that

$$\sin \theta = \frac{d}{2\sqrt{\frac{d^2}{4} + L^2}}$$

and

$$r^2 = \frac{d^2}{4} + L^2$$

we can write our field as

$$\begin{aligned}E_y &= \frac{2}{4\pi\epsilon_o} \frac{q}{(\frac{d^2}{4} + L^2)} \left(\frac{d}{2\sqrt{\frac{d^2}{4} + L^2}} \right) \\ &= \frac{1}{4\pi\epsilon_o} \frac{qd}{(\frac{d^2}{4} + L^2)^{\frac{3}{2}}}\end{aligned}$$

This is our total field at the distance L away on the axis. This is the environment that a mover charge could move through.

Note that we pretended that we had a mover, q_o , but in finding the field the q_o canceled out, so indeed we are left with just the environment in our calculation, we just have the field.

Now suppose our mover charge is very far away. That is, suppose we make L very large. So large that $L \gg d$ then

$$\lim_{L \gg d} \frac{1}{(\frac{d^2}{4} + L^2)^{\frac{3}{2}}} = \frac{1}{L^3}$$

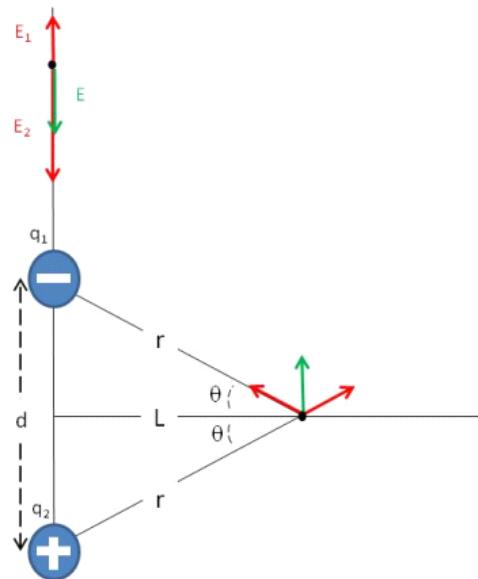
Then our field becomes

$$\begin{aligned} E &= E_y = \frac{2}{4\pi\epsilon_0} \frac{qd}{\left(\frac{d^2}{4} + L^2\right)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0} \frac{qd}{L^3} \end{aligned}$$

Since many charged particles are small, like atoms or molecules, this limit is often useful.

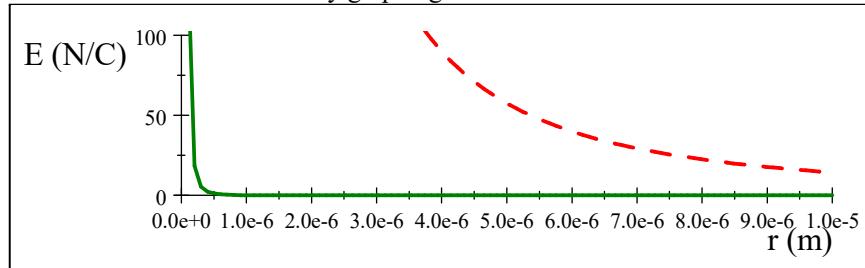
Suppose we repeat the calculation, but this time we chose a point that is L away, but that is on the y -axis above the charges, we would find

$$E = E_y = \frac{2}{4\pi\epsilon_0} \frac{qd}{L^3}$$



The result is similar, but the field is a little stronger in this direction.

Let's look at one of these cases by graphing it.



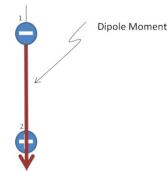
We can see that the dipole field (solid green line) falls off much faster than a point

charge field (dashed red line). This makes sense because the farther away we get, the more it looks like the two charges are right next to each other, and since they are opposite in sign, they are essentially neutral when viewed together from far away. We can see why atoms don't exhibit a significant charge forces at normal distances.

This arrangement of charges we already know as a dipole. We are treating the two charges as a unit making the environment in which other charges might move. Since we are treating the two charges as one unit, it is customary to define a quantity

$$p = qd$$

and to make this a vector by defining the direction of p to be from the negative to the positive charge along the axis.

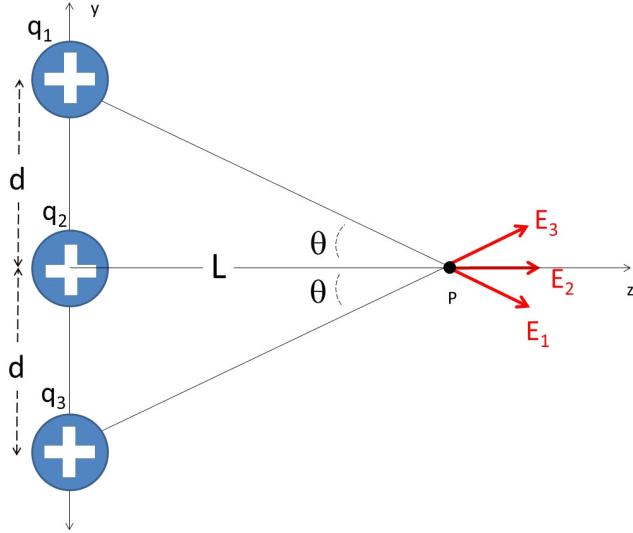


Then we can write the dipole field as

$$\vec{E}_y = \frac{2}{4\pi\epsilon_0} \frac{\vec{p}}{L^3}$$

We could also treat this dipole as a complicated mover charge in some other environmental field! Then this quantity \vec{p} will help us understand how a dipole will move when placed in an environmental electric field. For example, we know that water molecules are dipoles. A microwave oven creates a strong environmental electric field that makes the water molecules rotate. When we studied rotational motion we found a mass-like term that helped us to know how difficult something was to make rotate. That was the moment of inertia. This dipole term, \vec{p} , will tell us how likely a dipole is to spin, so we will call \vec{p} the *dipole moment*.

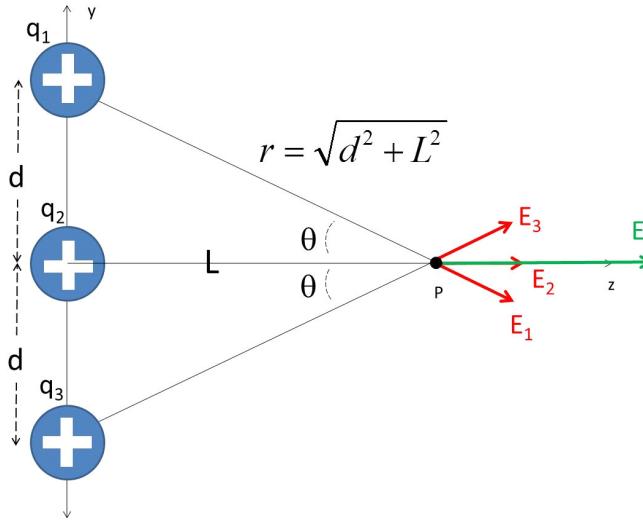
Three charges



Question 223.22.5

We are working our way toward many charges that will require using integration to sum up the contributions to the field. But let's make this transition slowly. Next let's add just one more environmental charge, for a total of three.

Let's just start with the fields this time. From our picture, we expect in this case to have only z -components. Since all the charges are the same sign,



then

$$E_{net_z} = E_1 \cos(-\theta) + E_2 + E_3 \cos(\theta)$$

We can guess from symmetry that

$$E_1 = E_3 = \frac{1}{4\pi\epsilon_o} \frac{q}{r^2}$$

But this time, since we have redefined d , the distance from q_1 and q_3 to the point P where we want to know the field is

$$r = \sqrt{d^2 + L^2}$$

so

$$E_1 = E_3 = \frac{1}{4\pi\epsilon_o} \frac{q}{(d^2 + L^2)}$$

and

$$E_2 = \frac{1}{4\pi\epsilon_o} \frac{q}{L^2}$$

and observing the triangles formed and remembering our trigonometry, we have

$$\cos \theta = \frac{L}{\sqrt{d^2 + L^2}}$$

so

$$\begin{aligned} E_z &= \frac{1}{4\pi\epsilon_o} \frac{q}{(d^2 + L^2)} \left(\frac{L}{\sqrt{d^2 + L^2}} \right) \\ &\quad + \frac{1}{4\pi\epsilon_o} \frac{q}{L^2} \\ &\quad + \frac{1}{4\pi\epsilon_o} \frac{q}{(d^2 + L^2)} \left(\frac{L}{\sqrt{d^2 + L^2}} \right) \end{aligned}$$

or

$$E_z = \frac{q}{4\pi\epsilon_o} \left(\frac{2L}{(d^2 + L^2)^{\frac{3}{2}}} + \frac{1}{L^2} \right)$$

This is our answer.

Once again let's consider the limit $L \gg d$. If our answer is right, when we get very far from the group of charges they should look like a single charge with the amount of charge being the sum of all three environmental charges. In this limit

$$\lim_{L \gg d} \frac{1}{(d^2 + L^2)^{\frac{3}{2}}} = \frac{1}{L^3}$$

so our limit becomes

$$\begin{aligned} E_z &\approx \frac{q}{4\pi\epsilon_o} \left(\frac{2L}{L^3} + \frac{1}{L^2} \right) \\ &= \frac{1}{4\pi\epsilon_o} \left(\frac{3q}{L^2} \right) \end{aligned}$$

so on the central axis

$$\vec{E} \approx \frac{1}{4\pi\epsilon_o} \left(\frac{3q}{L^2} \right) \hat{k}$$

And indeed, this is very like one charge that is three times as large as our actual charges if we get far enough away.

This shows us a pattern we will often see. Far away, our field looks like what we would

expect if the net charge were all congregated in a point. Near the charges, we must calculate the superposition of the fields. But far away we can treat the distribution as a point charge. This is very like what we did with mass in PH121 or Dynamics. We could often treat masses as point masses at the center of mass, if the distances involved were larger than the mass sizes.

Question 223.22.6

Of course, in these last two problems we picked nice places along axes to find the electric field. If we picked less convenient places we would have both y and z -components.

Combinations of many charges

We have found the field from a point charge.

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q_E}{r^2} \hat{r} \quad (22.1)$$

where the field is in the same direction as \hat{r} if the charge is positive, and in the opposite direction if the charge is negative (think of our field lines, they go toward the negative charge). This will become one of a group of standard charge configurations that we will use to gain a mental picture of complex configurations of charge. We have done this already for combinations of point charges. We can combine the point charge fields to get the total field.

The other standard models are combinations of many, many charges.

Line of Charge

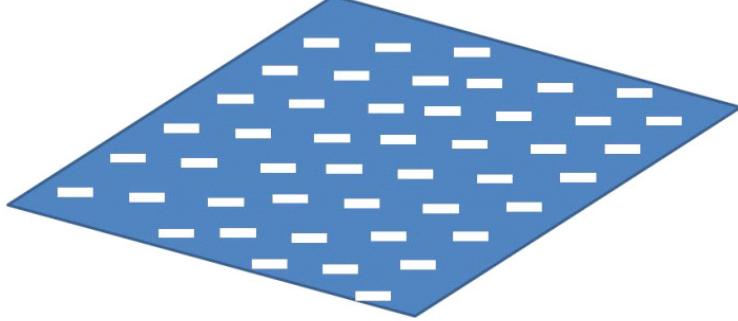
Another is an infinitely long line of charge, or a infinite charged wire. Since this long line of charge is infinite, it must have an infinite amount of charge. But we can describe “how much” charge it has with a linear charge density

$$\lambda = \frac{Q}{L}$$



Semi-infinite sheet of charge

A sheet or plane of charge, usually a semi-infinite sheet of charge is also useful



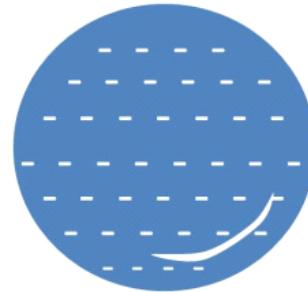
We have the same problem of having infinite charge, but if we define an amount of charge per unit area

$$\eta = \frac{Q}{A}$$

we can compare sheets that are more charge rich than others.

Sphere of charge

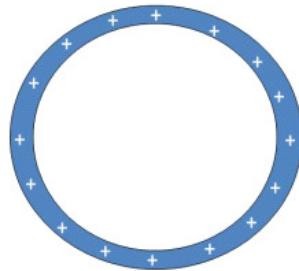
Finally, we have drawn a sphere of charge already



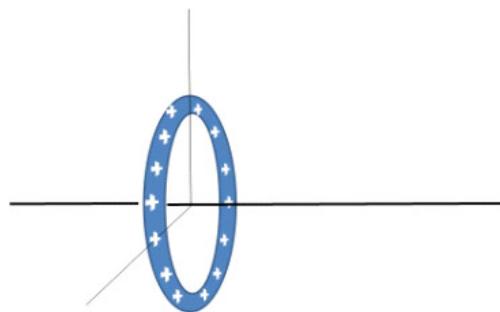
We can define an amount of charge per unit volume to help describe this distribution

$$\rho = \frac{Q}{V}$$

The spherical shell of charge is related to a sheet of charge, so we will include it here



This configuration of charge is drawn in cross section like the others. From your calculus experience you can guess that a spherical shell of charge with a certain volume charge density might be useful in integration, but we also can easily produce such a configuration of charge by charging a round balloon or a spherical conductor.



The ring of charge is similar to the spherical shell, but is also much like the line of charge.

In our next lecture, we will take on the job of finding the fields that result from these last few charge configurations except the spherical shell, which will have to wait a few lectures.

On-Line Visualizations

For a 2D visualization of the field try:

https://phet.colorado.edu/sims/html/charges-and-fields/latest/charges-and-fields_all.html

<https://icphysweb.z13.web.core.windows.net/simulation.html>

322 Chapter 22 Electric Fields of Standard Charge Configurations Part I

<http://www.falstad.com/emstatic/index.html>

And here is a 3D visualization:

<http://www.falstad.com/vector3de/>

Basic Equations

23 Electric Fields of Standard Charge Configurations Part II

Fundamental Concepts

- Integrating vector fields for continuous distributions of charge
 - Start with $\vec{E} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{r}$
 - Find an expression for dq
 - Use geometry to find expressions for r and to eliminate \hat{r}
 - Solve the integral

Fields from Continuous Charge Distributions

Question 223.23.1

Suppose we have a continuous distribution of charge with some mover charge q_m fairly far away. You might ask, how do we get a continuous distribution of charge? After all, charge seems to be quantized. Well, if we have a collection of charges where the distances between the individual charge carriers are much smaller than the distance from the whole collection of charges to some point where we want to measure the field (where the mover charge might be), then in our field calculations at this distant point we can model the charge distribution causing the field as continuous. As an analogy, think of your computer screen. It is really a collection of dots of light. But if we are a few feet away, we see a continuous picture. We can treat the dots as though there were no space in between them. For our continuous charge model, it is the same. We are supposing we are observing from far enough away that we won't notice the effects of the charges being separated by small distances.

We should remember, though, that this is a macroscopic view. At some point it must break down, since charge is carried in discrete amounts. If we want the field very close to a distribution of charges, we must treat our charge distribution as a collection of individual charges like we did in the last lecture. Notice in our last lecture that we found that the field infinitely far from the charges was always zero. That is too far away

for our continuous charge model to be useful. But if we went far enough away—but not too far, the three charge configuration looked like a point charge with a total charge that was the sum of the individual charges. At such distances, the separation between the charges become unimportant. This is the sort of large distance we are talking about in our continuous charge distribution model.

To find the field due to a continuous charge distribution, we break up the charged object into small pieces in a calculus sense. Each small piece is still a continuous distribution of charge. It will have an amount of charge Δq_E , where here the Δ means “a small amount of.” Then we calculate the field due to this element of charge. We repeat the process for each element using the superposition principle to sum up all the individual field contributions. This is very like our method of finding the field from individual charges, only instead of a sum we want to let Δq_E become very small and use an integral. The field due to this bit of charge is

$$\Delta \vec{E} = \frac{1}{4\pi\epsilon_0} \frac{\Delta q_E}{r^2} \hat{r}$$

Recall that here Δ means “a small bit of” and is not a difference between two charge values or two field values. We learned that we can sum up the fields from each piece

$$\begin{aligned} \vec{E} &\approx \sum_i \Delta \vec{E}_i \\ &\approx \frac{1}{4\pi\epsilon_0} \sum_i \frac{\Delta q_i}{r_i^2} \hat{r}_i \end{aligned}$$

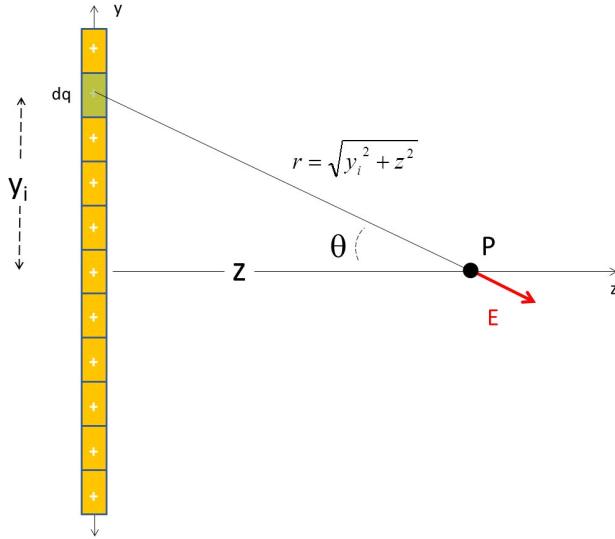
and now we use our M215 (or M113) tricks to convert this into an integral. We let our small element of charge become very small (but not so small that we violate our assumption that the charge distribution of Δq_E is continuous).

$$\begin{aligned} \vec{E} &= \lim_{\Delta q_i \rightarrow 0} \frac{1}{4\pi\epsilon_0} \sum_i \frac{\Delta q_i}{r_i^2} \hat{r}_i \\ &= \frac{1}{4\pi\epsilon_0} \int \frac{dq_E}{r^2} \hat{r} \end{aligned}$$

The limits of the integration must include the entire distribution of charge if we want the total field. This will be our basic equation for finding the field for continuous distributions of charge.

Let's do some examples.

Line of charge



Question 223.23.2

Let's try this for a line of charge. This may seem like a simple charge configuration, but this problem is really quite challenging. Let's say that the charge is evenly distributed along the line. Then we can use the linear charge density

$$\lambda = Q/L$$

to find dq . The quantity Q is the total amount of charge on the wire and L is the length of the wire. Then

$$dq = \lambda dy$$

Of course, we may not always have a constant density, then we need to have an element of charge that varies with position. For a line charge, we would have

$$dq = \lambda(y) dy$$

but for now, let's assume the linear charge density is constant. Our basic formula tells us that we should add up all the dq elements. But we have an obstacle. We need a different $\hat{\mathbf{r}}$ for every dq_i . How do we deal with this?

Just like with last lecture, we only need the component of the part of the field that does not cancel. Here we need to have drawn a good picture. From our drawing we can tell that, in this case, only the z component will survive (the y -components cancel). So we only need to find

$$E_z = \vec{\mathbf{E}} \cdot \hat{\mathbf{k}}$$

This is a good thing, because our basic equation has an $\hat{\mathbf{r}}$ in it

$$\vec{\mathbf{E}} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{\mathbf{r}}$$

and we don't know how to do this integral including the $\hat{\mathbf{r}}$. So we need to eliminate the direction part before we can proceed. Since we just need the z -component,

$$\begin{aligned} E_z &= \vec{\mathbf{E}} \cdot \hat{\mathbf{k}} \\ &= \frac{1}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{dq}{r^2} \hat{\mathbf{r}} \cdot \hat{\mathbf{k}} \end{aligned}$$

and we recognize

$$\hat{\mathbf{r}} \cdot \hat{\mathbf{k}} = \cos \theta$$

So we are left with just

$$E_z = \frac{1}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{dq}{r^2} \cos \theta$$

which is much more likely be integrable with what we know from M113 or M215.

Like in our last lecture, we will want to express

$$r = \sqrt{y^2 + z^2}$$

and it makes it easier if we write

$$\cos \theta = \frac{z}{\sqrt{y^2 + z^2}}$$

Then our integral can be written as

$$\begin{aligned} E_z &= \frac{1}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{\lambda dy}{y^2 + z^2} \frac{z}{\sqrt{y^2 + z^2}} \\ &= \frac{\lambda z}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{dy}{(y^2 + z^2)^{\frac{3}{2}}} \end{aligned}$$

This now looks like a M215 or M113 problem. We can find this integral in an integral table or you can use your calculator, or a symbolic math package, or you can remember your M215 or M113 and prove that

$$\int_{-L/2}^{L/2} \frac{dx}{(x^2 \pm a^2)^{\frac{3}{2}}} = \frac{\pm x}{a^2 \sqrt{x^2 \pm a^2}}$$

so

$$\begin{aligned}
 E_z &= \frac{\lambda z}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{dy}{(y^2 + z^2)^{\frac{3}{2}}} \\
 &= \frac{\lambda z}{4\pi\epsilon_0} \left[\frac{y}{z^2 \sqrt{y^2 + z^2}} \right]_{-L/2}^{L/2} \\
 &= \frac{\lambda z}{4\pi\epsilon_0} \left[\frac{L/2}{z^2 \sqrt{(L/2)^2 + z^2}} - \frac{-L/2}{z^2 \sqrt{(-L/2)^2 + z^2}} \right] \\
 &= \frac{\lambda}{4\pi z \epsilon_0} \frac{L}{\sqrt{(L/2)^2 + z^2}} \\
 &= \frac{1}{4\pi\epsilon_0} \frac{Q}{z \sqrt{\left(\frac{L}{2}\right)^2 + z^2}}
 \end{aligned}$$

This is the field due to a charged rod of length L .

Note that there are only a few integrals that we can solve in closed form to find electric fields. It might be a good idea to build your own integral table for our exams, including the integrals from the problems and examples we work.

An infinitely long line of charge is one of our basic charge models. So far our line of charge is not infinitely long. We can find the field due to an infinite line of charge by letting L become large

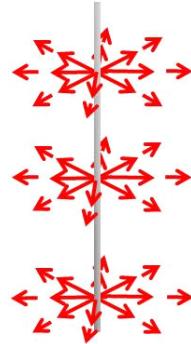
$$\begin{aligned}
 E_z &= \lim_{L \rightarrow \infty} \frac{1}{4\pi\epsilon_0} \frac{Q}{z \sqrt{\left(\frac{L}{2}\right)^2 + z^2}} \\
 &= \frac{1}{4\pi\epsilon_0} \frac{Q}{z \left(\frac{L}{2}\right)} \\
 &= \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z} \\
 &= \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z}
 \end{aligned}$$

or if we use r now in place of z to define the distance from the center of the line of charge (so it is easier to compare to our point charge formula), we have

$$\vec{E}_z = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{r} \hat{k}$$

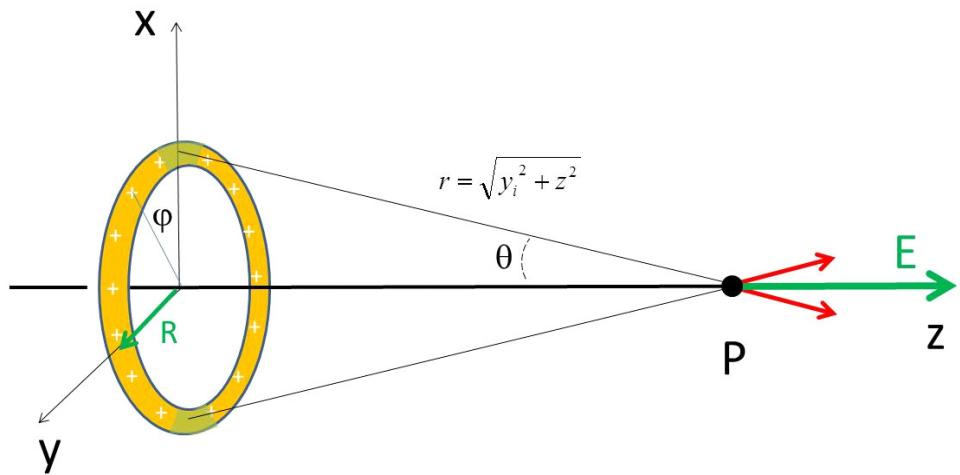
Question 223.23.3

We should get a mental picture of what this means.



The field around a long line of charge only depends on the distance away from the line, and on the linear charge density. As we would expect, the field gets weaker as we get farther away. But it does not get weaker as fast as the point charge case. That makes some sense, because our infinite line of charge is, well, really big. You are never really too far away from something that is infinitely big. So we should not expect such a charge configuration to look very like a point charge no matter how far away we go. Of course an infinite line of charge is not something we can really build. So this is a useful approximation near, say, a charged wire. But farther from the wire the approximation would not be so good and we would have to go back to our finite line solution.

Ring of charge



Question 223.23.4

Using what we have learned from the line of charge, we can find the axial field of a ring of charge. Again, our picture is critically important. We will need to solve the problem of eliminating $\hat{\mathbf{r}}$. From the picture, we can see that we will only have a z -component again. So we can eliminate $\hat{\mathbf{r}}$ the same way as in the last problem. We model the ring as a line of charge of length $2\pi R$ that has been bent into a circle. Again we have the basic equation

$$\vec{\mathbf{E}} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{\mathbf{r}}$$

Since the ring of charge is like a line of charge bent into a hoop. So we can plan to work this problem very like the the line charge. Start again with

$$dq = \lambda dy$$

but now we know that for the hoop

$$dq = \lambda ds$$

where s is the arc length. Recall that

$$\begin{aligned} s &= R\phi \\ ds &= Rd\phi \end{aligned}$$

where R is the radius of the ring and ϕ is an angle measured from the x -axis. So our dq expression becomes

$$dq = \lambda R d\phi$$

For the whole ring

$$\begin{aligned} Q &= \lambda R 2\pi \\ &= 2\pi R \lambda \end{aligned}$$

We also need to use geometry to find r , the distance to our point were we want to know the field.

$$r = \sqrt{y_i^2 + z^2}$$

but since this is a ring, our $y_i = R$ for all i . So

$$r = \sqrt{R^2 + z^2}$$

and using the same reasoning as in our last problem,

$$\cos \theta = \frac{z}{\sqrt{R^2 + z^2}}$$

Then we can set up our integral.

$$\begin{aligned} E_z &= \vec{\mathbf{E}} \cdot \hat{\mathbf{k}} \\ &= \frac{1}{4\pi\epsilon_0} \int_{-L/2}^{L/2} \frac{dq}{r^2} \hat{\mathbf{r}} \cdot \hat{\mathbf{k}} \end{aligned}$$

Putting in all the parts we have found yields

$$\begin{aligned} E_z &= \frac{1}{4\pi\epsilon_0} \int \frac{dq}{R^2 + z^2} \frac{z}{\sqrt{R^2 + z^2}} \\ &= \frac{1}{4\pi\epsilon_0} \int \frac{z\lambda R d\phi}{(R^2 + z^2)^{\frac{3}{2}}} \\ &= \frac{z\lambda R}{4\pi\epsilon_0 (R^2 + z^2)^{\frac{3}{2}}} \int_0^{2\pi} d\phi \end{aligned}$$

This is an easy integral to do! and we see that the axial field is

$$E_z = \frac{z2\pi R\lambda}{4\pi\epsilon_0 (R^2 + z^2)^{\frac{3}{2}}}$$

or, using our form for Q

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{zQ}{(R^2 + z^2)^{\frac{3}{2}}} \hat{k}$$

Once again we should check to see if this is a reasonable result. If we take the limit as z goes to infinity, we get zero. That is comforting. But if we just let z be much larger than R , but not too big

$$\begin{aligned} \lim_{z \gg R} \vec{E} &= \lim_{z \gg R} \frac{1}{4\pi\epsilon_0} \frac{zQ}{(R^2 + z^2)^{\frac{3}{2}}} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \frac{zQ}{(z^2)^{\frac{3}{2}}} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \frac{zQ}{z^3} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \frac{Q}{z^2} \hat{k} \end{aligned}$$

we again have a point charge field with total charge Q . Since a ring of charge should look like a point charge if we get far enough away, this is reasonable.

We have worked two problems for continuous charge distributions. The pattern for solving both problems was the same. And we will follow the same pattern for solving for the field from continuous charge distributions in all our problems:

- Start with $\vec{E} = \frac{1}{4\pi\epsilon_0} \int \frac{dq_E}{r^2} \hat{r}$
- Find an expression for dq_E
- Use geometry to find an expressions for r , the distance from dq_E to the point, P , where we want to know the field.
- Eliminate \hat{r}
- Solve the integral

If you have a harder problem, one where you need the field from a continuous charge distribution at a point that is not on an axis, or your problem has little symmetry, you

can go back to

$$\begin{aligned}\vec{\mathbf{E}} &\approx \sum_i \Delta \vec{\mathbf{E}}_i \\ &\approx \frac{1}{4\pi\epsilon_o} \sum_i \frac{\Delta q_i}{r_i^2} \hat{\mathbf{r}}_i\end{aligned}$$

and perform the sum numerically. We won't do this in our class, but you might in practice or in a higher level electrodynamics course.

Basic Equations

The basic equation from this chapter is the equation for finding the field from a distribution of charge

$$\vec{\mathbf{E}} = \frac{1}{4\pi\epsilon_o} \int \frac{dq_E}{r^2} \hat{\mathbf{r}}$$

The process for using this equation is

- Start with $\vec{\mathbf{E}} = \frac{1}{4\pi\epsilon_o} \int \frac{dq_E}{r^2} \hat{\mathbf{r}}$
- Find an expression for dq
- Use geometry to find an expression for r
- Eliminate $\hat{\mathbf{r}}$ in the usual way by turning a two or three-dimensional problem into two or three one-dimensional problems (using vector components, etc.)
- Solve the integral(s) (Don't forget to report the direction)

24 Motion of Charged Particles in Electric Fields

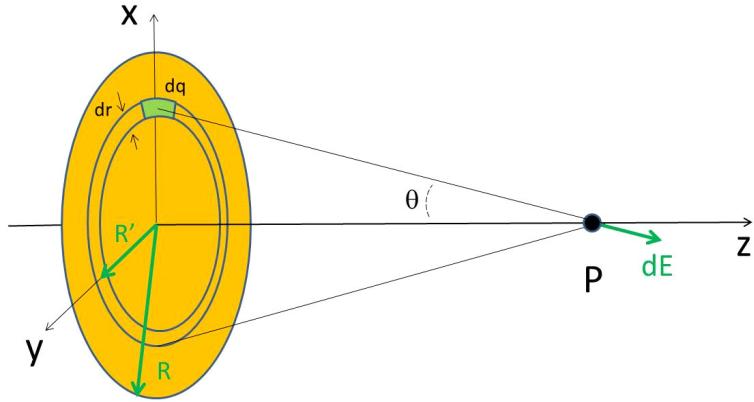
Fundamental Concepts

- The capacitor
- Field of an ideal Capacitor
- Motion of particles in a constant electric field

Sheet of Charge

Question 223.23.5

Let's try a two dimensional distribution of charge, a uniform flat sheet of charge. We will assume that the sheet is infinitely large (so we don't have to deal with what happens at the edges). Let's call the surface charge density $\eta = Q/A$ where Q is the total charge and A is the total area. Of course, we can't calculate this surface charge density directly from the totals, because they are infinite. But we could take a square meter of area and find the amount of charge in that small area. The ratio should be the same for any area so long as η is uniform. We will find the electric field to the right of the sheet at point P .



Once again we start with

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{r}$$

We need to find dq , an expression for r , and get rid of \hat{r}

Since the disk is uniformly charged, then, knowing the surface charge density

$$\eta = \frac{Q}{A}$$

we can find the total amount of charge for an area

$$Q = \eta A$$

so

$$dq = \eta dA$$

but what area, dA , should we use? Notice the green patch in the figure that is marked dA . Think for a moment about arc length

$$s = R\phi$$

This little area is about $ds = Rd\phi$ long, and about dR wide. If we let dA be small enough, this is exact. So

$$dA = Rd\phi dR$$

Then our dq is just η times this

$$dq = \eta Rd\phi dR$$

From geometry we identify

$$r = \sqrt{z^2 + R^2}$$

and, due to symmetry we expect only the z -component of the field to survive. So to get rid of \hat{r} we multiply (dot product) by \hat{k} . There will be an angle, θ , between \hat{r} and \hat{k} . So

we expect the result of the dot product to be that we multiply by the cosine of θ

$$\cos \theta = \frac{z}{\sqrt{z^2 + R^2}}$$

We want to put all this into our basic equation. This time the radius R changes, so let's call it R' so we recognize that it is a variable over which we must integrate, then

$$\vec{\mathbf{E}} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{\mathbf{r}}$$

becomes

$$\begin{aligned} E_z &= \frac{1}{4\pi\epsilon_0} \int \frac{\eta R' d\phi dR'}{(z^2 + R'^2)} \hat{\mathbf{r}} \cdot \hat{\mathbf{k}} \\ &= \frac{1}{4\pi\epsilon_0} \int \frac{\eta R' d\phi dR'}{(z^2 + R'^2)} \cos \theta \end{aligned}$$

and we will integrate from $R' = 0$ to $R' = R$.

$$E_z = \frac{1}{4\pi\epsilon_0} \int_0^{2\pi} \int_0^R \frac{z\eta R' d\phi dR'}{(z^2 + R'^2)^{\frac{3}{2}}}$$

Performing the integration over $d\phi$ just gives us a factor of 2π

$$E_z = \frac{z\eta\pi}{4\pi\epsilon_0} \int_0^R \frac{2R' dR'}{(z^2 + R'^2)^{\frac{3}{2}}}$$

where, for convenience, we have left the 2 inside the integral (it will be useful later).

We need to solve the integral over dR' . A u -substitution is one way. Suppose we let

$$u = z^2 + R'^2$$

so

$$du = 2R' dR'$$

We will need to adjust the limits of integration, for $R' = 0$ we have

$$u = z^2$$

and for $R' = R$

$$u = z^2 + R^2$$

then our integral becomes

$$E_z = \frac{z\pi\eta}{4\pi\epsilon_0} \int_{z^2}^{z^2 + R^2} \frac{du}{(u)^{\frac{3}{2}}}$$

We get

$$\begin{aligned}
 E_z &= \frac{z\pi\eta}{4\pi\epsilon_o} \left[\frac{-2}{(u)^{\frac{1}{2}}} \right]_{z^2}^{z^2+R^2} \\
 &= \frac{z\pi\eta}{4\pi\epsilon_o} \left(\frac{-2}{(z^2+R^2)^{\frac{1}{2}}} - \frac{-2}{(z^2)^{\frac{1}{2}}} \right) \\
 &= \frac{-2z\pi\eta}{4\pi\epsilon_o} \left(\frac{1}{(z^2+R^2)^{\frac{1}{2}}} - \frac{1}{z} \right) \\
 &= \frac{-2\pi\eta}{4\pi\epsilon_o} \left(\frac{z}{(z^2+R^2)^{\frac{1}{2}}} - 1 \right) \\
 &= \frac{-2\pi\eta}{4\pi\epsilon_o} \left(\frac{1}{\frac{1}{z}(z^2+R^2)^{\frac{1}{2}}} - 1 \right)
 \end{aligned}$$

The result is

$$E_z = \frac{2\pi\eta}{4\pi\epsilon_o} \left(1 - \frac{1}{\left(1 + \frac{R^2}{z^2}\right)^{\frac{1}{2}}} \right)$$

or

$$E_z = \frac{2\pi\eta}{4\pi\epsilon_o} \left(1 - \left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \right)$$

This looks messy, but this is the answer.

But wait, this is really a disk of charge with radius R . We wanted an infinite sheet of charge. So. suppose we let R get very big. Then

$$\begin{aligned}
 E_{R \rightarrow \infty} &= \lim_{R \rightarrow \infty} \frac{2\pi\eta}{4\pi\epsilon_o} \left(1 - \left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_o} \\
 &= \frac{\eta}{2\epsilon_o}
 \end{aligned}$$

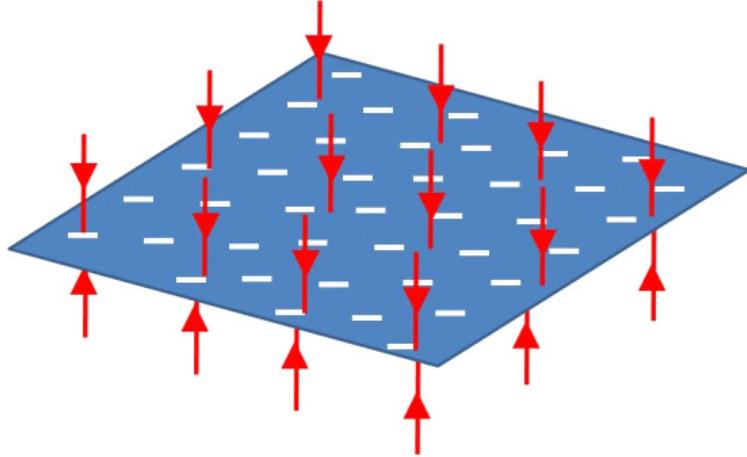
Question 223.23.6

This is the field for our semi-infinite sheet of charge.

We should take some time to figure out if this makes sense.

This sheet cuts the entire universe into to two parts. It is so big, that it is hard to say anything is very far away from it. So we can understand this answer, The field from such a sheet of charge is constant every where in all of space. No matter how far away we get, it will never look like a point charge, in fact, it never really looks any farther away!

Note we did just one side of the sheet, there is a matching field on the other side. So this sheet of charge fills all of space with a constant field.



Of course this is not physically possible to build, but we will see that if we look at a large sheet of charge, like the plate of a capacitor, that near the center, the field approaches this limit, because the sides of the sheet are far away.

Visualization
falstad 3D

Let's go back and consider the disk of charge.

$$E_z = \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \right)$$

Suppose we look at this distribution from very far away for a finite disk. We expect that it should look like a point charge with total charge Q . Let's show that this is true. When z gets very large R/z is very small.

$$E_{z \gg R} = \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \right)$$

Let's look at just the part

$$\left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}}$$

This is of the form $(1 + x)^n$ where x is a small number. We can use the binomial expansion

$$(1 + x)^n \approx 1 + nx \quad x \ll 1$$

to write this as

$$\left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \approx 1 - \frac{1}{2} \frac{R^2}{z^2}$$

so in the limit that z is large we have

$$\begin{aligned} E_{z \gg R} &= \frac{2\pi\eta}{4\pi\epsilon_o} \left(1 - 1 + \frac{1}{2} \frac{R^2}{z^2} \right) \\ &= \frac{1}{4\pi\epsilon_o} \frac{\pi \frac{Q}{\pi R^2} R^2}{z^2} \\ &= \frac{1}{4\pi\epsilon_o} \frac{Q}{z^2} \end{aligned}$$

Which looks like a point charge as we expected. We have just a small, disk of charge very far away. That is looks like a point charge with total charge Q .

Spheres, shells, and other geometries.

I won't do the problem for the field of a charged sphere or spherical shell. We could, but we will save them for a new technique for finding fields from configurations of charge that we will learn soon. This new technique will attempt to make the integration much easier.

Constant electric fields

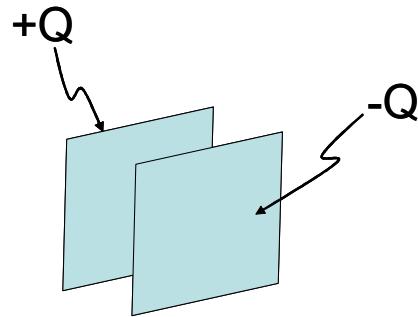
Let's try to use what we know about electric fields to predict the motion of charged particles that are placed in electric fields. We will start with the simplest case, a charged particle moving in a constant electric field. Before we take on such a case, we should think about how we could produce a constant electric field.

We know that a semi-infinite sheet of charge produces a constant electric field. But we realize that a semi-infinite object is hard to build and hard to manage. But if the size of the sheet of charge is very large compared to the charge size, using our solution for a semi-infinite case might not be too bad if we are away from the edges of the real sheet.

We want to study just such a device. In fact we will use two finite sheets of charge.

Capacitors

From what we know about charge and conductors, we can charge a large metal plate by touching it to something that is charged, like a rubber rod, or a glass rod that has been rubbed with the right material.



If we have two large metal plates and touch one with a rubber rod and one with a glass rod, we get two oppositely charged sheets of charge.

What would the field look like for this oppositely charged set of plates? Here is one of our thread-in-oil pictures of just such a situation. We are looking at the plates edge-on.

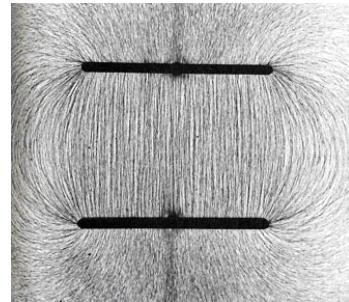
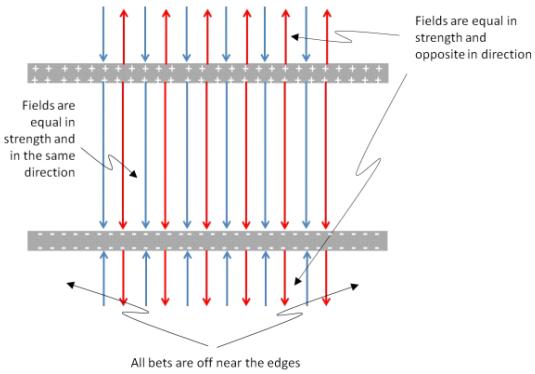


Figure 24.16.http://stargazers.gsfc.nasa.gov/images/geospace_images/electricity/charged_plates.jpg

Near the center, the field is close to constant. Near the sides it is not so much so. We are probably justified in saying the field in the middle is nearly constant. A look at the field lines shows us why



Note that in between the plates, the electric field from the positive plate is downward. But so is the electric field from the negative plate. The two fields will add together. Outside the plates, the field from one plate is in the opposite direction from that of the other plate. The two fields will nearly cancel. If our device is made of semi-infinite sheets of charge, they will precisely cancel, because the field of a semi-infinite sheet of charge is uniform everywhere.

We call this configuration of two charged plates a *capacitor* and, as you might guess, this type of device proves to be more useful than just making nearly constant fields. It is a major component in electronic devices. Before we can build an iPad or a laptop, we will need to understand several different types of basic devices. This set of charged plates is our first.

Of course, for real capacitors, the fields outside cancel completely only near the center of the plates. Near the edges, the direction of the fields will change, and we get the sort of behavior that we see in figure 24.16 near the edges.

It is probably worth noting that outside the capacitor the field has a *magnitude of zero* (or nearly zero). It is not really correct to say that there is no field. In fact, there are two superimposed fields, or alternately, a field from each of the charges on each plates, all superimposed. The fields are there, but their magnitude is zero.

In the middle, then, we will have

$$\begin{aligned} E &= E_+ + E_- \\ &\approx \frac{\eta}{2\epsilon_o} + \frac{\eta}{2\epsilon_o} \\ &= \frac{\eta}{\epsilon_o} \\ &= \frac{Q}{A\epsilon_o} \end{aligned}$$

Particle motion in a uniform field

Question 223.24.1

Now that we have a way to form a uniform electric field, we can study charged particles moving in this field. Motion of particles in uniform fields is really something we are familiar with. It is very much like a ball in a uniform gravitational field. But we have the complication of having two different types of charge. The force on such a particle is given by

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

but we can combine this with Newton's second law

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

to find the particle's acceleration

$$\vec{a} = \frac{q_m \vec{E}}{m}$$

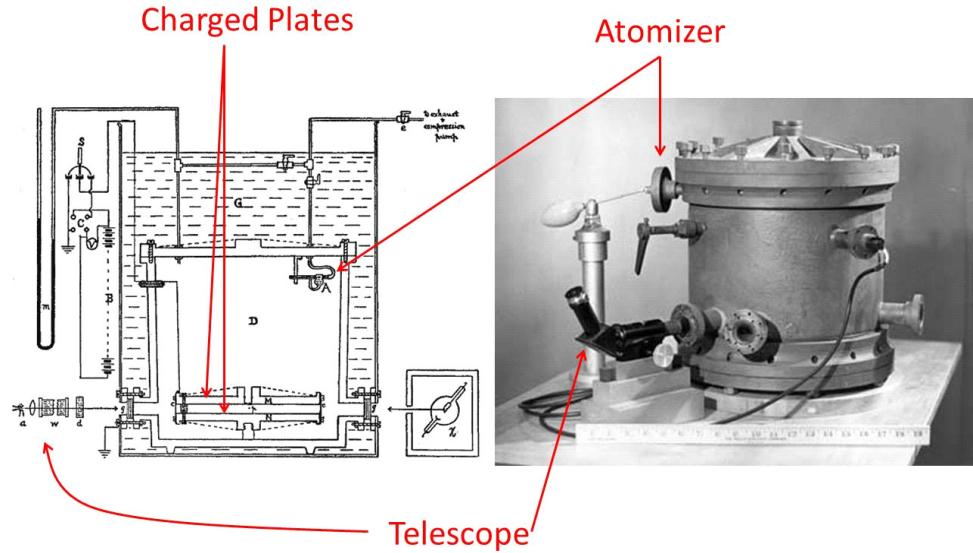
Note, this is NOT true in general. It is only true for constant electric fields.

CRT Demo

Millikan

Let's try a problem. Perhaps you have wondered, "how do we know that charge comes in packets of the size of the electron charge?" This is a good story, and uses many of the laws we have learned.

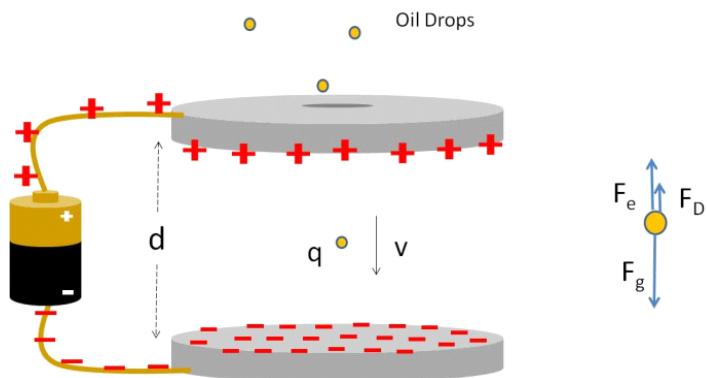
Robert Millikan devised a clever device in the early 1900's. A picture of his device is given below.



Millikan's oil-drop apparatus: Diagram taken from orginal Millikan's paper, 1913,

Image taken in 1906 (Both Images in the Public Domain)

Schematically we can draw the experiment like this.



Millikan made negatively charged oil drops with an atomizer (fine spray squirt bottle). The drops are introduced between two charged plates into what we know is essentially a constant electric field. A light shines off the oil drops, so you can see them through a telescope (not shown). We can determine the motion of the oil drops just like we did in PH121 or Dynamics. If the upper plate has the positive charge, then the electric field \vec{E} is downward. A free body diagram for a drop is shown in the figure to the left of the apparatus. We can write out Newton's second law for the drop (our mover charge).

$$\Sigma F_y = m_d a_y = -F_g \pm F_D + F_e$$

where F_D is a drag force because we have air resistance.

If the upper plate has the positive charge, then the electric field $\tilde{\mathbf{E}}$ is downward. So

$$\vec{\mathbf{F}}_e = -q_m \vec{\mathbf{E}}$$

The field points down, the charge is negative, so the force is upward (positive in our favorite coordinate system). We can write newton's second law as

$$m_d a_y = -F_g \pm F_D + q_m E$$

If F_e is large enough, we can make the oil drop float up! Then the drag force is downward

$$m_d a_y = -mg - F_D + q_m E$$

and if we are very careful, we can balance these forces so we have the drop float upward at a small constant velocity.

$$0 = -mg - F_D + q_m E$$

The constant speed is really slow, hundredths of a centimeter per second. so we can watch the drop move with no problem (except for patience). Once he achieved a constant speed, by knowing the drop size and density Millikan could calculate the mass, and therefore the charge.

$$mg + F_D = q_m E$$

we see that

$$q_m = \frac{mg + F_D}{E}$$

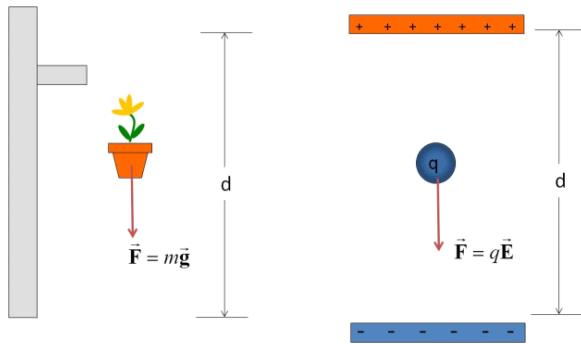
Which is where our problem ends. But Millikan went farther. He had actual data, so he could compare charges on different droplets. He found that no matter what the value for q_m , it was a multiple of a value, $q_e = 1.602 \times 10^{-19} \text{ C}$. So

$$q_m = nq_e \quad n = 0, \pm 1, \pm 2, \dots$$

to within about 1%.¹⁹ So the smallest charge the drops could have added to them was $1 \times q_e$ and any other larger charge would be a larger multiple of q_e . The conclusion is that charge comes in units of q_e . We recognize q_e as the electron charge. You can't add half of an electron charge. This experiment showed that charge seems to only comes in whole units!

Free moving particles

¹⁹ There is actually some controversy about this. Apparently Millikan and his students threw out much of their data, keeping only data on drops that behaved like they thought they should. They were lucky that this poor analysis technique did not lead to invalid results! (William Broad and Nicholas Wade, *Betrayers of the truth*, Simon and Schuster, 1983)



We may recall that for an object falling in a gravitational field, say, near the Earth's surface, the acceleration, g , is nearly constant. If we have a charge moving in a constant electric field, we have a constant acceleration. From Newtons' second law,

$$ma = q_m E$$

we can see that this acceleration is

$$a = \frac{q_m E}{m}$$

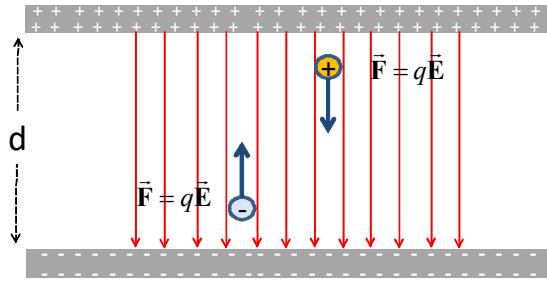
From our Dynamics or PH121 experience, we have a set of equations to handle problems that involve constant acceleration

$$\begin{aligned} x_f &= x_i + v_{ix}\Delta t + \frac{1}{2}a_x\Delta t^2 \\ v_{xf} &= v_{xi} + a_x\Delta t \\ v_{xf}^2 &= v_{xi}^2 + 2a_x\Delta x \end{aligned}$$

and

$$\begin{aligned} y_f &= y_i + v_{iy}\Delta t + \frac{1}{2}a_y\Delta t^2 \\ v_{yf} &= v_{yi} + a_y\Delta t \\ v_{yf}^2 &= v_{yi}^2 + 2a_y\Delta y \end{aligned}$$

These are known as the *kinematic equations*. You derived them if you took Dynamics (or derived them and then memorized them if you took PH121). Let's try a brief problem. Suppose we have a positive charge in a uniform electric field as shown.



Let $y = 0$ at the positive plate. How fast will the particle be going as it strikes the negative plate?

We use the acceleration

$$\begin{aligned} a_y &= \frac{q_m E}{m} \\ a_x &= 0 \end{aligned}$$

For this problem we don't have any x -motion, So we can limit ourselves to.

$$\begin{aligned} y_f &= y_i + v_{iy}\Delta t + \frac{1}{2} \left(\frac{q_m E}{m} \right) \Delta t^2 \\ v_{yf} &= v_{yi} + \left(\frac{q_m E}{m} \right) \Delta t \\ v_{yf}^2 &= v_{yi}^2 + 2 \left(\frac{q_m E}{m} \right) \Delta y \end{aligned}$$

We don't have the time of flight of the particle, but we can identify

$$\Delta y = d$$

The particle started from rest, so

$$v_{yi} = 0$$

Therefore it makes sense to use the last of the three equations, because we know everything that shows up in this equation but the final speed, and that is what we want to find.

$$\begin{aligned} v_{yf}^2 &= v_{yi}^2 + 2 \left(\frac{q_m E}{m} \right) \Delta y \\ v_{yf}^2 &= 0 + 2 \left(\frac{q_m E}{m} \right) d \\ v_{yf} &= \sqrt{\frac{2q_m Ed}{m}} \end{aligned}$$

There is a complication, however. With gravity, we only have one kind of mass. But with charge we have two kinds of charge. Suppose we have a negative particle.

Of course the negative particle would not move if it was started from the positive side. It would be attracted to the positive plate. But suppose we start the negative particle from the negative plate. It would travel “up” to the positive plate. We defined the downward direction as the positive y -direction without really thinking about it. Now we realize that the upward direction must be opposite, so upward is the negative y -direction. Our negative particle will experience a displacement $\Delta y = -d$.

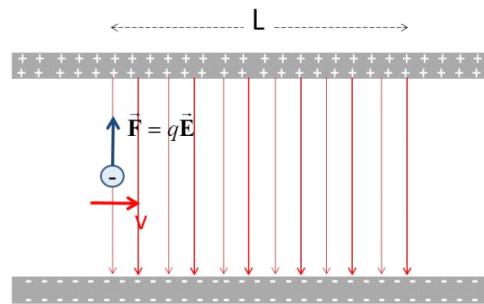
Then

$$\begin{aligned} v_{yf}^2 &= v_{yi}^2 + 2 \left(\frac{-q_m E}{m} \right) \Delta y \\ v_{yf}^2 &= 0 + 2 \left(\frac{-q_m E}{m} \right) (-d) \\ v_{yf} &= \sqrt{\frac{2q_m Ed}{m}} \end{aligned}$$

we get the same speed, but this illustrates that we will have to be careful to watch our signs.

In this last problem we have had only an electric force, no gravitational force. This is important to notice. If there were also a gravitational force, we would need to use Newton’s second law to add up the forces like we did with the Millikan problem.

Let’s take another example. This time let’s send in a negatively charged particle horizontally through the capacitor. The particle will move up due to the electric field force. How far up will it go as it travels across the center of the capacitor?



Let’s define the starting position as

$$\begin{aligned} x_i &= 0 \\ y_i &= 0 \end{aligned}$$

We can identify that

$$\begin{aligned} v_{ix} &= v_0 \\ v_{iy} &= 0 \end{aligned}$$

And that

$$\begin{aligned} a_y &= \frac{q_m E}{m} \\ a_x &= 0 \end{aligned}$$

We can fill in these values in our kinematic equations

$$\begin{aligned} x_f &= 0 + v_{xi} t + \frac{1}{2} (0) \Delta t^2 \\ v_{fx} &= v_{ix} + (0) t \\ v_{xf}^2 &= v_{ix}^2 + 2 (0) \Delta x \end{aligned}$$

and

$$\begin{aligned} y_f &= 0 + (0) \Delta t + \frac{1}{2} \left(\frac{q_m E}{m} \right) \Delta t^2 \\ v_{yf} &= (0) + \left(\frac{q_m E}{m} \right) \Delta t \\ v_{yf}^2 &= (0) + 2 \left(\frac{q_m E}{m} \right) (y_f - 0) \end{aligned}$$

From the first set we see that $v_{fx} = v_{ix}$, that is, the x -direction velocity does not change. That makes sense because we have no force component in the x -direction.

After t seconds we see that the charged particle has traveled a distance

$$x_f = v_{xi} t$$

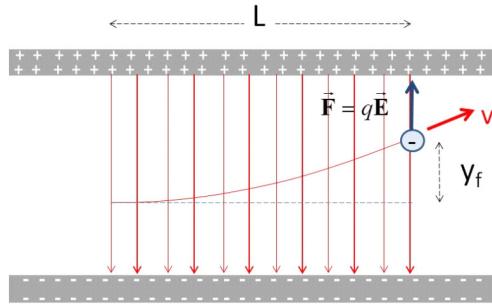
If we measure $x_f = L$ then we can see how long it took for the particle to travel through the capacitor

$$t = \frac{L}{v_{ix}}$$

Now let's look at the deflection. We can use the first equation of the y -set

$$\begin{aligned} y_f &= \frac{1}{2} \left(\frac{q_m E}{m} \right) t^2 \\ &= \frac{1}{2} \left(\frac{q_m E}{m} \right) \left(\frac{L}{v_{ix}} \right)^2 \end{aligned}$$

Let's see if this makes sense. If the electric field gets larger, the particle will deflect more.



This is right. The field causes the force, so more field gives more effect from the force. If we increase the charge, the deflection grows since the force depends on the charge of the moving particle. This also seems reasonable. If the mass increases, it is harder to move the particle, so it makes sense that a larger mass makes a smaller deflection. If the particle is in the field longer, the deflection will increase, so the dependence on L makes sense. Finally, if the initial speed is larger the particle spends less time in the field, so the deflection will be less.

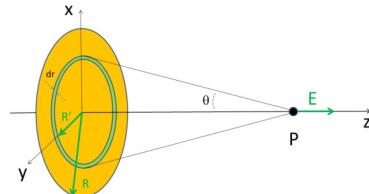
Of course all of this depends on the field being uniform. For a non uniform field the force is still

$$\vec{F} = q_m \vec{E}(x, y, z)$$

but now the field is a function of position. This makes for a more difficult problem. For now we will stick to constant fields. If we had to take on a non-uniform field, we would likely use a numerical technique.

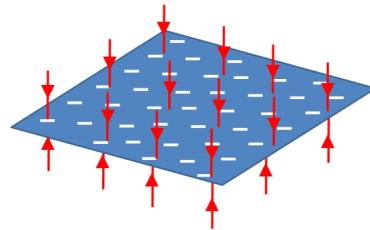
Basic Equations

The magnitude of the electric field due to a disk of charge along the disk's axis



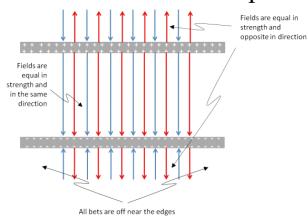
$$E_z = \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \left(1 + \frac{R^2}{z^2} \right)^{-\frac{1}{2}} \right)$$

The magnitude of the electric field due to a semi-infinite sheet of charge



$$E = \frac{\eta}{2\epsilon_0}$$

The magnitude of the electric field inside an ideal capacitor



$$E = \frac{Q}{A\epsilon_0}$$

Motion of a charged particle in a constant electric field

$$\vec{a} = \frac{q_m \vec{E}}{m}$$

$$\begin{aligned} x_f &= x_i + v_{ix}t + \frac{1}{2}a_x t^2 & y_f &= y_i + v_{iy}t + \frac{1}{2}a_y t^2 \\ v_{xf} &= v_{xi} + a_x t & v_{yf} &= v_{yi} + a_y t \\ v_{xf}^2 &= v_{xi}^2 + 2a_x \Delta x & v_{yf}^2 &= v_{yi}^2 + 2a_y \Delta y \end{aligned}$$

25 Dipole motion, Symmetry

Fundamental Concepts

- Force and torque on a dipole in a uniform field
- Force on a dipole in a non-uniform field
- Drawing the shape of a field using symmetry

This lecture combines two topics that might be better separated. The first relates to forces on charges in uniform fields. This is what we discussed last lecture. The next is the beginning of the ideas that will allow us to use symmetry and geometry to avoid integration over charges. But because our lecture times are only an hour, and we can only do so much at once, they are combined here together. But they form a nice transition between the two topics this way. We will first study the motion of dipoles in uniform, and not so uniform fields. We will find symmetry and geometry plays a part in our solutions. Then we will study the fields of standard symmetric objects.

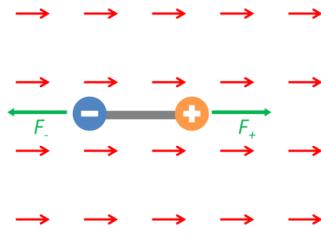
Dipole motion in an electromagnetic field

We remember dipoles, a pair of charges of equal magnitude but opposite in charge, bound together at set separation distance. Let's take our environment to be a constant electric field, and our mover to be a dipole.

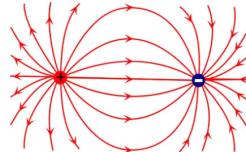


Question 223.25.1

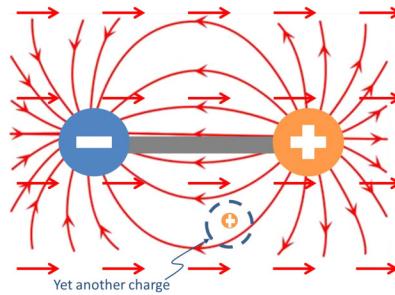
Here is a diagram of the situation.



Notice that as usual, just the environmental field is drawn. There is a field from the dipole, too,

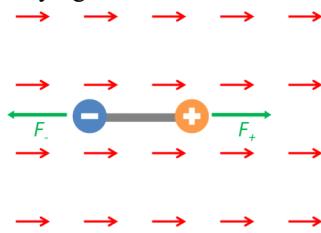


but this is the mover's self-field and it cannot create a force on the dipole, so we will not draw it. Of course, if we introduce yet another charge, q_{new} , the environmental field this new charge would feel would be a combination of both the dipole field and the uniform field! We would have to draw the superposition of the two fields.

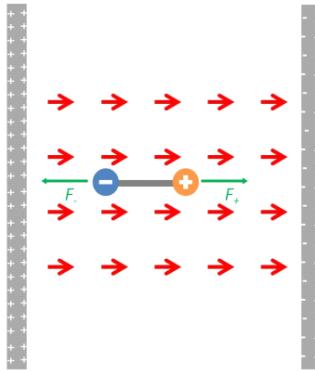


But that is a different problem!

Here is our case again. We only draw the environmental field that will cause the motion of the mover object we are studying.



To understand these figures, we have to remember that the red field arrows are an *external field* that is, the dipole is not making this field, so something else must be. We did not draw that something else. Since it is a uniform field, it is probably a capacitor. Here is what it might look like



The positive side must be to the left, because the red external field arrows come from the left. The negative side must be to the right, because the field arrows are pointed that direction. We can get away with not drawing the source of the external field because the force on the dipole charges is just

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

If we know \vec{E} , then we don't need any information about its source to find the force. Since the field is the environment that the mover charges feel, the field is enough. Let's find the net force on the dipole due to the environmental field.

Question 223.25.2

We use Newton's second law to find that

$$F_{net_x} = -F_{-E} + F_{+E} = ma_x$$

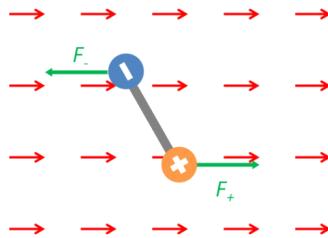
and our definition of the electric field to find

$$\begin{aligned} F_{-E} &= q_- E \\ F_{+E} &= q_+ E \end{aligned}$$

so, since $|q_-| = |q_+| = q$

$$-qE + qE = ma_x$$

which tells us that there is no acceleration, no net force. The center of mass of a dipole does not accelerate in a uniform field. But we remember from PH121 that we can make things rotate.



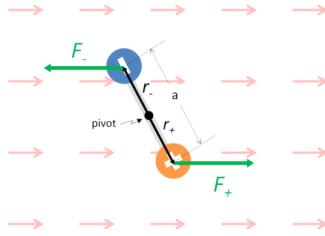
Question 223.25.3

If the dipole is not aligned with its axis in the field direction, then the forces will cause a torque (or moment).

We remember that torque is given by

$$\vec{\tau} = \vec{r} \times \vec{F}$$

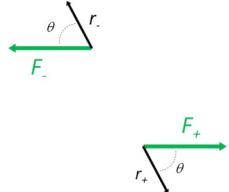
substituting in our force and defining the distance between the charges to be a we can write this out



The magnitude of the torque is given by

$$|\tau| = rF \sin \theta_{rF}$$

where θ_{rF} is the angle between \mathbf{r} and \mathbf{F} . It is easier to find that angle if we redraw each displacement vector from the pivot and each force with their tails together



Then for one charge, say, q_-

$$\tau = \frac{a}{2}qE \sin \theta$$

We use the right-hand-rule that you learned in Dynamics or PH121 to find the direction. We can see that the direction will be out of the page. But we have two charges, so we have a torque from each charge. A quick check with the right-hand-rule for torques will convince us that the direction for the torque due to q_+ is also out of the page, and the magnitude is the same, so our total torque is

$$\begin{aligned} \tau_{net} &= \tau_+ + \tau_- \\ &= aqE \sin \theta \end{aligned}$$

Question 223.25.4

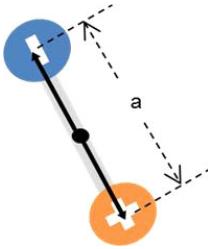
which we can write as

$$\tau_{net} = pE \sin \theta$$

or the *dipole moment*, p , multiplied by $E \sin \theta$. Recalling the form of a cross product

$$\vec{A} \times \vec{B} = AB \sin \theta \hat{n}$$

where \hat{n} is perpendicular to both \vec{A} and \vec{B} , we have a hint that we could write our torque as a cross product. We would have to make p a vector, though. So let's define \vec{p} as a vector with magnitude aq and make its direction along the line connecting the charge centers, with the direction from negative to positive.



Then we can write the torque as

$$\vec{\tau} = \vec{p} \times \vec{E} \quad (25.1)$$

which is our form for the torque on a dipole.

Let's try a problem. Let's find the maximum angular acceleration for a dipole.

Recall that Newton's second law for rotational motion is

$$\Sigma\tau = I\alpha$$

where I is the moment of inertia and α is the angular acceleration. Then we can find how the dipole will accelerate

$$\alpha = \frac{\tau_{net}}{I}$$

For a dipole, I is simple

$$\begin{aligned} I &= m_- r_-^2 + m_+ r_+^2 \\ &= m \left(\frac{a}{2}\right)^2 + m \left(\frac{a}{2}\right)^2 \\ &= \frac{1}{2}ma^2 \end{aligned}$$

so our acceleration is

$$\begin{aligned} \alpha &= \frac{pE \sin \theta}{\frac{1}{2}ma^2} \\ &= \frac{2pE \sin \theta}{ma^2} \end{aligned}$$

Suppose we look at this for a water molecule in a microwave oven. What is the maximum angular acceleration experienced by the water molecule if the oven has a field strength of $E = 200 \text{ V/m}$?

The dipole moment for a water molecule is something like

$$p_w = 6.2 \times 10^{-30} \text{ C m}$$

and the separation between the charge centers is something like

$$a = 3.9 \times 10^{-12} \text{ m}$$

and the molecular mass of water is

$$M = 18 \frac{\text{g}}{\text{mol}}$$

which is

$$M = mN_A$$

so the mass of a water molecule is

$$\begin{aligned} m &= \frac{M}{N_A} = \frac{18 \frac{\text{g}}{\text{mol}}}{6.022 \times 10^{23} \frac{1}{\text{mol}}} \\ &= 2.989 \times 10^{-26} \text{ kg} \end{aligned}$$

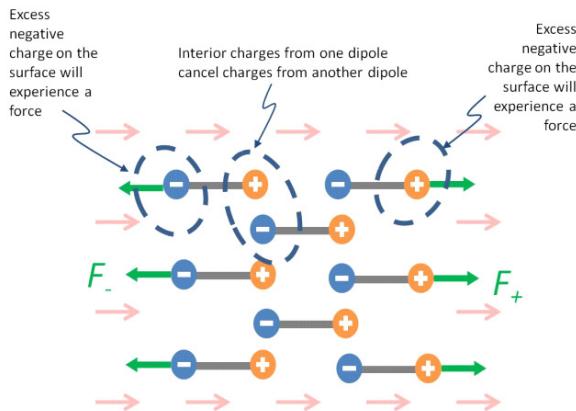
then when $\sin \theta = 1$ we will have a maximum

$$\begin{aligned} \alpha &= \frac{2(6.2 \times 10^{-30} \text{ C m})(200 \text{ V/m})}{(2.989 \times 10^{-26} \text{ kg})(3.9 \times 10^{-12} \text{ m})^2} \\ &= 5.455 \times 10^{21} \frac{\text{rad}}{\text{s}^2} \end{aligned}$$

Our numbers were kind of rough estimates, but still the result is amazing. Imagine if this happened inside of you! which is why we really should be careful with microwave ovens and microwave equipment.

Induced dipoles

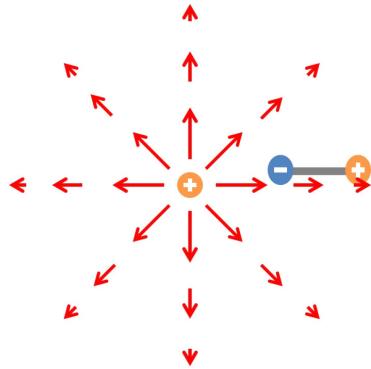
Suppose that we place a large insulator in a uniform electric field.



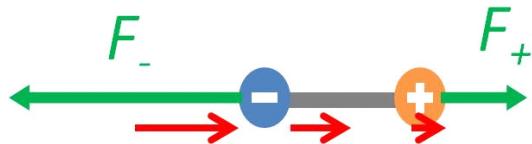
The atoms tend to polarize and become dipoles. We say we have *induced* dipoles within the material. Notice that in the middle of the insulator there is still no net charge. But because we have made the atoms into dipoles, one side of the insulator becomes negatively charged and the other side becomes positively charged. This does not create a net force, but we will find that separating the charges like this can be useful in building capacitors.

Non-uniform fields and dipoles

Suppose we place our dipole in a non-uniform field. Of course the result will depend on the field, so let's take an example. Let's place a dipole in the field due to a point charge.



We can see that the field is much weaker at the location of the positive charge than it is at the negative charge location. If we zoom in on the location near our dipole we can see that now we will have an acceleration!



$$\Sigma F_x = -F_{-E} + F_{+E} = ma_x$$

so

$$-qE_{\text{large}} + qE_{\text{small}} = ma_x$$

Let's go back to our charged balloon from many lectures ago. We found that the charge "leaked off" our balloon. We can see why now. The water molecules in the air are attracted to the charges, and stick to them. When the water molecules float off, they will take our charge with them. For this problem, the dipole is the environment and our balloon electron is the moving charge. We can calculate the net force easily with our field from a dipole that we found earlier,

$$\vec{E}_y = \frac{2}{4\pi\epsilon_o} \frac{\vec{p}}{L^3}$$

then the force on the electron on the balloon is

$$\begin{aligned} F &= q_e E \\ &= \frac{2q_e}{4\pi\epsilon_0} \frac{p}{L^3} \end{aligned}$$

So if the dipole is about a 0.01 cm away

$$\begin{aligned} F &= \frac{2(1.602 \times 10^{-19})}{4\pi(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2})} \frac{6.2 \times 10^{-30} \text{ C m}}{(0.01 \text{ cm})^3} \\ &= 1.7862 \times 10^{-26} \text{ N} \end{aligned}$$

But wait! we used the dipole as the environmental object and the single charge as the mover. So this is the force on the single charge! But by Newton's third law, the force on the dipole due to the electron must have the same magnitude and opposite direction so

$$F_{dipole} = -1.7862 \times 10^{-26} \text{ N}$$

We could do this problem the other way, thinking of the point charge as the environment and the dipole as the moving object. We know Coulomb's law for a point charge. So we use it to find the force on the individual parts of the dipole. We have to be careful because the minus charge is at a different r value than the positive charge.

$$\begin{aligned} -qE_- + qE_+ &= ma_x \\ -q\left(\frac{1}{4\pi\epsilon_0}\frac{Q}{r_-^2}\right) + q\left(\frac{1}{4\pi\epsilon_0}\frac{Q}{r_+^2}\right) &= ma_x \end{aligned}$$

or

$$\frac{Qq}{4\pi\epsilon_0} \left(\frac{1}{r_+^2} - \frac{1}{r_-^2} \right) = ma_x = F_{net}$$

this is the net force on a dipole due to the point charge.

The effective charge on one side of the water molecule is

$$\begin{aligned} q &= \frac{p}{a} = \frac{6.2 \times 10^{-30} \text{ C m}}{3.9 \times 10^{-12} \text{ m}} \\ &= 1.5897 \times 10^{-18} \text{ A s} \end{aligned}$$

(how can this be true?) so if the dipole is about a 0.01 cm away then

$$\begin{aligned} F_{net} &= \frac{(1.0 \times 10^{-19})(1.5897 \times 10^{-18} \text{ A s})}{4\pi(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2})} \\ &\quad \times \left(\frac{1}{\left(0.01 \text{ cm} + \frac{3.9 \times 10^{-12} \text{ m}}{2}\right)^2} - \frac{1}{\left(0.01 \text{ cm} - \frac{3.9 \times 10^{-12} \text{ m}}{2}\right)^2} \right) \\ &= -1.1150 \times 10^{-26} \text{ N} \end{aligned}$$

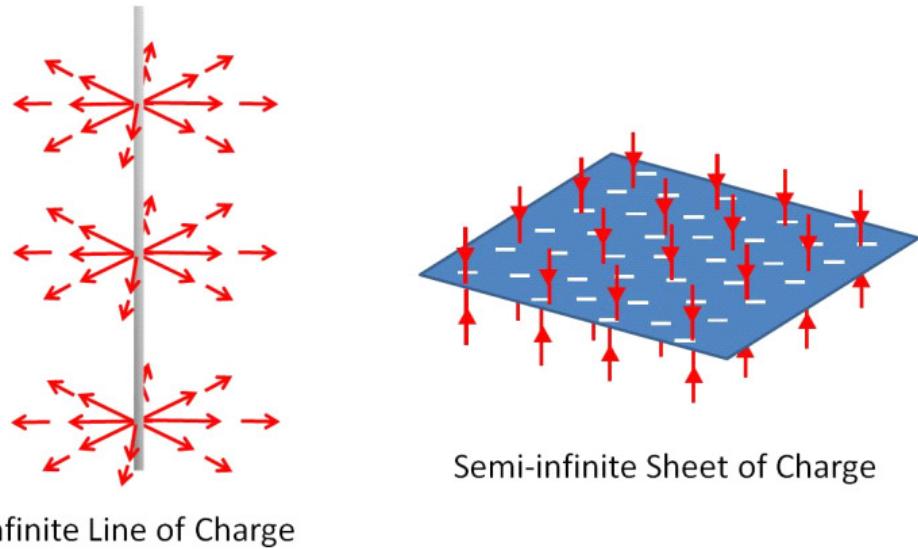
We expect the negative sign, both forces should be to the left. The answers are different, but within one order of magnitude. This is pretty good since for our dipole field we assumed that the distance from the dipole is very large and 0.01 cm is a somewhat

shorter version of very large!

Symmetry

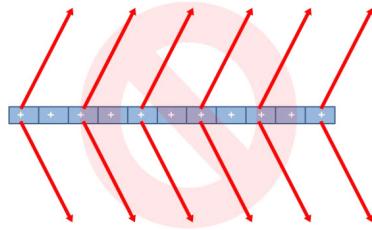
The symmetry of the uniform field figured strongly in the dipole problem. When the shape of the field changed, so did the resulting motion. This suggests that we could solve some problems just knowing the symmetry, or at least that symmetry might help us do simple predictions to help get a problem started. We need to be able to predict the field lines of a geometry to draw a picture to start solving a problem.

We have run into two geometries so far that have been helpful

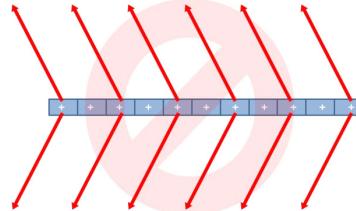


The infinite line of charge and the semi-infinite sheet of charge. We have found for the sheet that the field is constant everywhere. This is strongly symmetric. We could envision translating the sheet within the plane right or left. The field would look the same. We could envision reflecting the sheet so the left side is now the right side. That would also not change the field. We can say that the field of the sheet would be symmetric about translation within the plane of the sheet and symmetric on reflection.

Suppose we look at the sheet side-on. Suppose that we thought the field came off the sheet at an angle as shown.



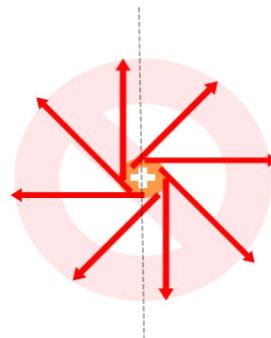
Notice that if we shift the sheet right or left, the field would still look the same, but if we reflected the sheet about the y -axis. Then we would have



But (and here is the important part) the shape of the charge distribution did not change on reflection. The sheet really looks just the same. It does not make sense that we should change the shape of the field if the shape of the charge distribution did not change. So we can tell that this can't be the right field shape.

Question 223.25.5

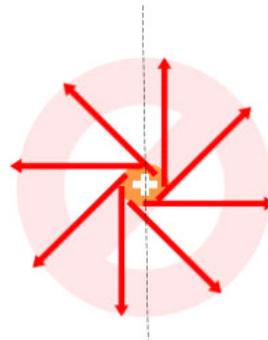
We can do this with any symmetric distribution of charge. Think of the infinite line of charge. If we move it left or right the field definitely changes. So it is not symmetric about translation along, say, the x -axis. But if we move the wire along its own axis, (for my coordinate system, along the y -axis) it should be symmetric because the charge distribution won't look different. We can guess from the last example that the field must come straight out perpendicular to the line of charge. It must be perpendicular, but what direction? Look at this end view. The field lines do come straight out, so this meets our criteria for being perpendicular to the line.



Line of Charge, End View

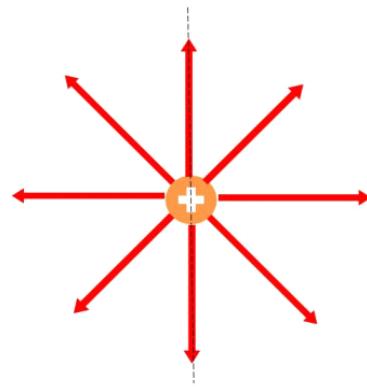
We could rotate the line about the axis of the line. Then the charge distribution would

look just the same. The field would also look just the same on rotation. But if we reflect the charge distribution across the axis shown, the charge distribution looks just the same, but the field would change.



Line of Charge, Reflection

We can tell that this is not the right field. We can tell that the field should look more like this.

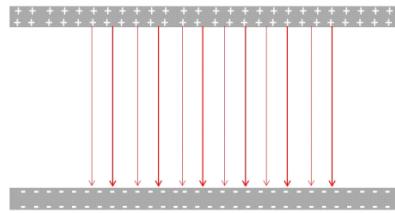


Line of Charge, End View

Combinations of symmetric charge distributions

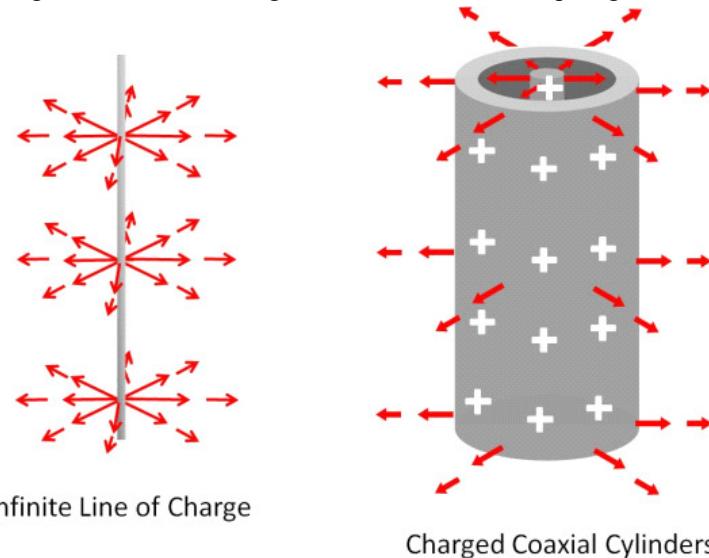
Question 223.25.6

We can combine sheets or lines of charge to build more complex systems. We did this to form a capacitor



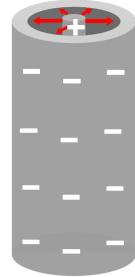
The field lines follow our symmetry guidelines. Because of the symmetry of the sheet of the field lines must be perpendicular to the sheets.

Again building from the line of charge, we can build more complex geometries



In the figure we have two positively charged concentric cylinders. The field is very reminiscent of a line charge field, and we can see that it must be using the same symmetry rules.

Of course the cylinders don't have to have the same charge.

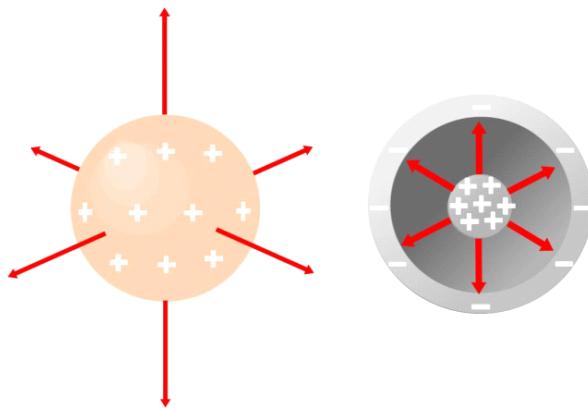


Oppositely Charged Coaxial
Cylinders

If the interior cylinder is positively charged and the exterior cylinder is negatively charged, we have a situation much like the capacitor. Each cylinder has a field outside the system, but those fields cancel out if there are equal charges on each cylinder. This situation is similar to a coaxial cable, and we will revisit it later in the course.²⁰

For the charge configurations we have drawn so far, we must keep in mind that they are infinite in at least one dimension. Finite configurations of charge in lines or sheets will have curved fields at the ends. The fields will be symmetric on reflection about their centers, but not on translation of any sort. Still, we will continue to use semi-infinite approximations in this class, and these constructs are good mental images under many circumstances.

Of course we can have a sphere. Spheres are very symmetrical, so we can guess using our symmetry ideas that the field from a charged sphere should be perpendicular to the surface of the sphere everywhere.



²⁰ Indeed, this coaxial cables have a capacitance!

We can see that this is true for both the sphere and for concentric spheres or any configuration of charge that is spherical.

Basic Equations

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

26 Electric Flux

Fundamental Concepts

- Electric flux is the amount of electric field that penetrates an area.
- An area vector is a vector normal to the area surface with a magnitude equal to the area.
- For closed surfaces, flux going in is negative and flux going out is positive by convention.

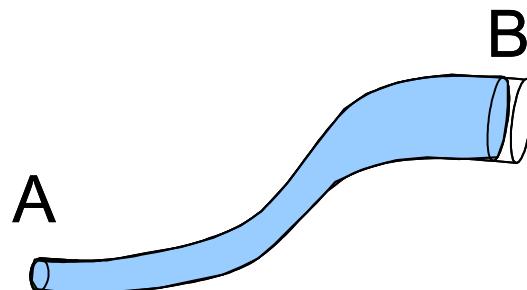
The Idea of Flux

Van de Graaff Generator Demo

Question 223.26.1

If you took PH123 or have had a class that deals with fluids, I can use an analogy (if not, you will probably be OK, because you have probably used a garden hose). Let's recall some fluid dynamics for a moment. Remember what we called a *flow rate*? This was from the equation of continuity

$$v_1 A_1 = v_2 A_2$$



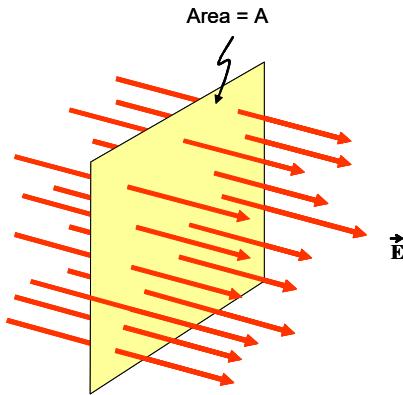
We wanted to know how much liquid was going by a particular part of the pipe in a given unit of time. We called vA a *flow rate*.

The idea of electric flux

I want to introduce an analogous concept. But this time I want to use the electric field instead of water speed

$$\Phi = EA$$

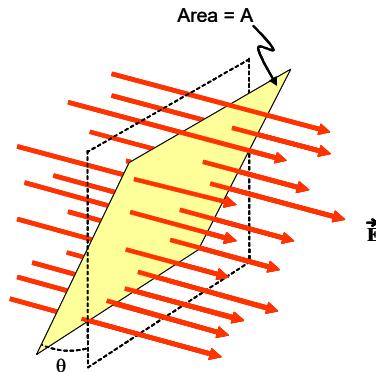
This is just like our flow rate in some ways. It is something multiplied by an area. In fact, it is how much of something goes through an area. We could guess that it is the amount of electric field that passes through the area, A . Now the electric fields we have dealt with so far don't flow. They just stay put (we will let them change later in the course). So it is only *like* a flow rate. But it is useful to think of this as "how much of something passes by an area," and the "something" is the electric field in this case. Let's consider a picture



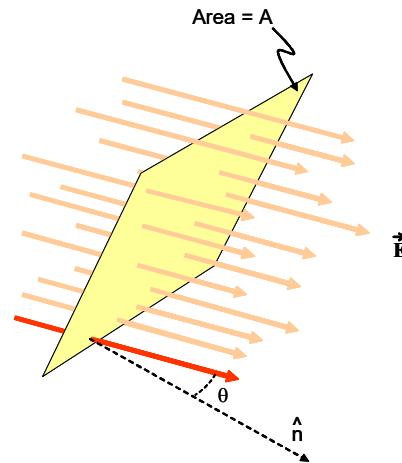
In this picture, we have a rectangular area, A , and the red arrows represent the field lines of the electric field. We can picture the quantity, Φ , as the number of field lines that pass through A . Remember that the number of field lines we draw is greater if the field strength is higher, so this quantity, Φ , tells us something about the strength of the field over the area.

Question 223.26.2

But, what if the area, A , is not perpendicular to the field?



We define an angle, θ (our favorite greek letter, but we could of course use β or α , or ζ or whatever) that is the angle between the field direction and the area. A more mathematical way to do this is to define a vector that is perpendicular to (normal to) the surface \hat{n} . Then we can use this vector and one of the field lines to define θ . It will be the angle between \hat{n} and the field lines.



Of course either way gives the same θ .

Now our definition of Φ can be made to work. We want the number of field lines passing through A , but of course, now there are fewer lines passing through the area because it is tilted. We can find Φ using θ as

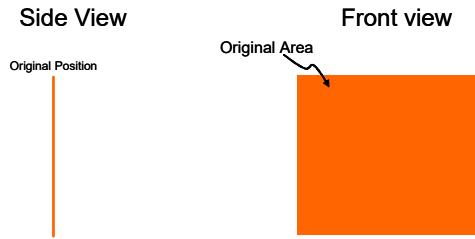
$$\Phi = EA \cos \theta \quad (26.1)$$

but let's consider what

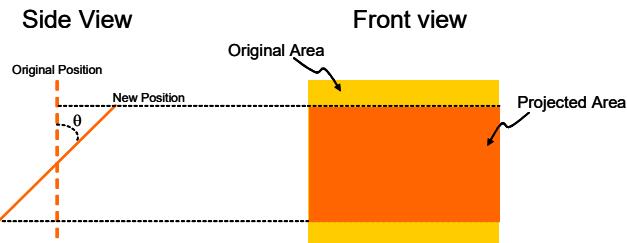
$$A \cos \theta$$

Tip a flat object

means. We can start with our original area.



If we tip the area, it looks smaller



The smaller area is called the *projected area*.

We can see that by tipping our area, we get fewer field lines that penetrate that area.

Really the number of field lines is just proportional to E , so we won't ever really count field lines. But this is a good mental picture for what flux means. Really we will calculate

$$\Phi = EA \cos \theta$$

The $\cos \theta$ with two magnitudes (field strength and area) multiplying it should remind you of something. It looks like the result of a vector dot product. If E and A were both vectors, then we could write the flux as

$$\Phi = \vec{E} \cdot \vec{A} \quad (26.2)$$

Domenstrate with a document with writing on one side

Well, we can define a vector that has A as its magnitude and is in the right direction to make

$$\vec{E} \cdot \vec{A} = EA \cos \theta$$

We define the *area vector*

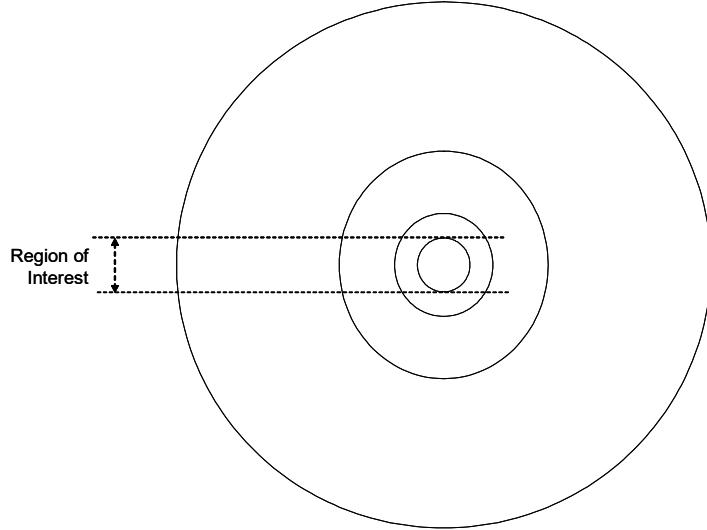
$$\vec{A} = \hat{n} A \quad (26.3)$$

Notice that for an open surface (one that does not form a closed surface with a empty space inside) we have to choose which side \hat{n} will point from. We can choose either side. But once we have made the choice, we have to stick with it for the entire problem we are solving.

Flux and Curved Areas

Trifold paper

Suppose the area we have is not flat? Then what? Well let's recall that if we take a sphere the surface will be curved. But if we take a bigger sphere, and look at the same amount of area on that sphere, it looks less curved.



This becomes more apparent if we remove the rest of the circle or sphere to take away the visual cues our eyes and minds use to say something is curved



Suppose we take a curved surface but we just look at a very small part of that surface. This would be very like magnifying our circle. We would see an increasingly flat surface piece compared to our increased scale of our image.

This gives us the idea that for an element of area, ΔA we could find an element of flux $\Delta\Phi$ for this small part of the whole curved surface. Essentially ΔA is flat (or we would just take a smaller ΔA).

$$\Delta\Phi = \vec{E} \cdot \vec{\Delta A} \quad (26.4)$$

This is just a small piece of the total flux through the curved surface, the total flux

through our whole curved surface is

$$\Phi_E \approx \sum \Delta\Phi \quad (26.5)$$

Of course, to make this exact, we will take the limit as $\Delta A \rightarrow 0$ resulting in an integral.

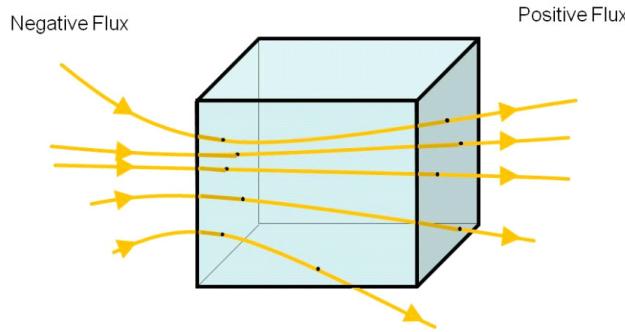
We find the flux through a curved surface to be

$$\Phi_E = \lim_{\Delta A \rightarrow 0} \sum_i \vec{E} \cdot \Delta \vec{A}_i = \int_{surface} \vec{E} \cdot d\vec{A} \quad (26.6)$$

Notice that this is a *surface integral*. It may be that you have not done surface integrals for some time, but we will practice in the upcoming lectures.

Closed surfaces

Suppose we build a box with our areas.



Then we would have some lines going in and some going out. By convention we will call the flux formed by the ones going in negative and the flux formed by the ones going out positive. From these questions we see that if there is no charge inside of the box, the net flux must be zero. We could take any size or shape of closed surface and this would be true! But if we do have charge inside of the box we expect there to be a net flux. If it is a negative net charge, it will be an negative flux and if it is a positive net charge it will be a positive net flux. Next lecture we will formalize this as a new law of physics, but for now we need to remember from M215 or M113 how to write an integration over a closed surface. We use a special integral sign with a circle

$$\Phi_E = \oint \vec{E} \cdot d\vec{A} \quad (26.7)$$

You will also see this written as

$$\Phi_E = \oint E_n dA \quad (26.8)$$

where E_n is the component of the field normal to the surface at the point area increment

Question 223.26.3
Required

Question 223.26.4
Required

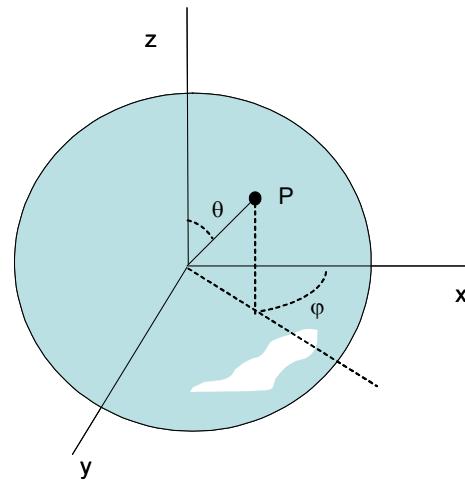
Question 223.26.5

dA .

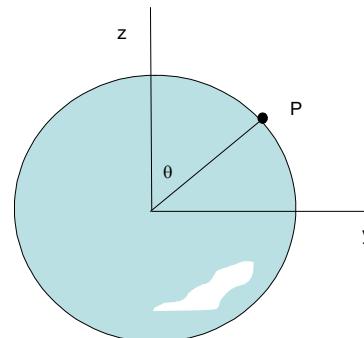
Flux example: a sphere

For each type of surface we choose, we need an area element to perform the integration. This is a lot like finding dq in our electric field integral. Let's take an example, a sphere.

We can start by finding the coordinates of a point, P , on the surface of the sphere.

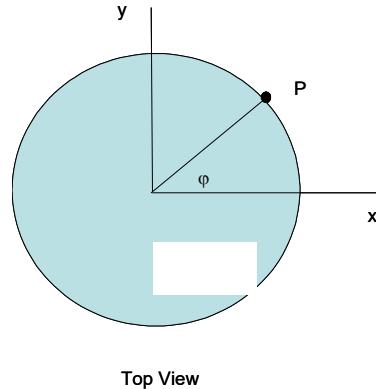


We define the coordinates in terms of two angles, θ and ϕ . Let's look at them one at a time. First θ



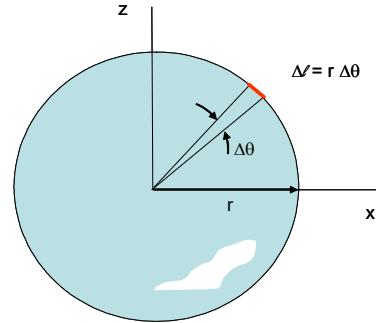
Side View

and now ϕ



Top View

Let's build an area by defining a sort of box shape on the surface by allowing a change in θ and ϕ ($\Delta\theta$ and $\Delta\phi$). First $\Delta\theta$,



Side View

The angle θ just defines a circle that passes through the “north pole” and “south pole” of our sphere. By changing θ we get a small bit of arc length. We remember that the length of an arc is

$$s_\theta = r\theta \quad (26.9)$$

where θ is in radians. So we expect that

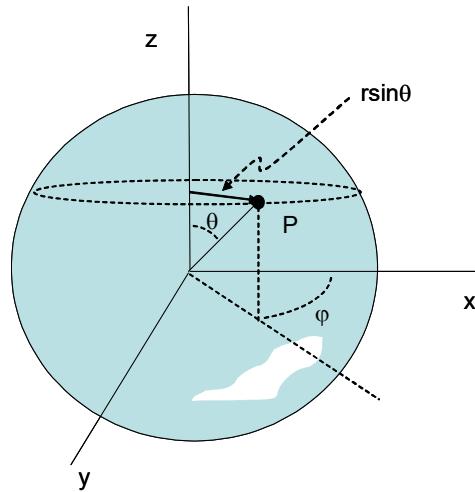
$$\Delta s_\theta = r\Delta\theta \quad (26.10)$$

We can check this by integrating

$$\int_0^{2\pi} r d\theta = r \int_0^{2\pi} d\theta = 2\pi r \quad (26.11)$$

Just as we expect, the integral of arc length around the whole circle is the circumference of the circle. Then Δs_θ is one side of our small box-like area, the box height.

Now let's look at ϕ

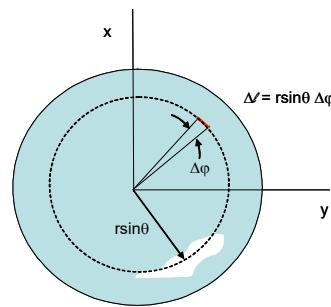


ϕ also forms a circle on the sphere, but its size depends on θ . Near the north pole, the radius of the ϕ -circle is very small. At $\theta = 90^\circ$, the ϕ -circle is in the xy plane and has radius r . We can write the radius of the ϕ -circle as a projection over $90^\circ - \theta$ which gives us a radius of $r \sin \theta$. Then we use the arc length formula again to find

$$s_\phi = (r \sin \theta) \phi \quad (26.12)$$

a change in arch length will be

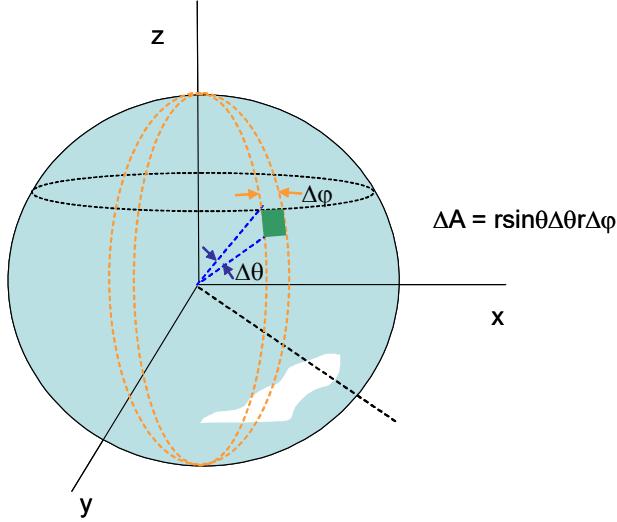
$$\Delta s_\phi = (r \sin \theta) \Delta \phi \quad (26.13)$$



Top View

This is the other side of our box, the box width.

Now let's combine them. We multiply $\Delta s_\theta \times \Delta s_\phi$ to obtain a roughly rectangular area.



$$\Delta A \approx \Delta s_\theta \times \Delta s_\phi = r \Delta \theta r \sin \theta \Delta \phi \quad (26.14)$$

which is the area of our small box. We have found an element of area on the surface of the sphere! Let's check our element of area by integration. After changing Δ to d and rearranging

$$dA = r^2 \sin \theta d\theta d\phi \quad (26.15)$$

then

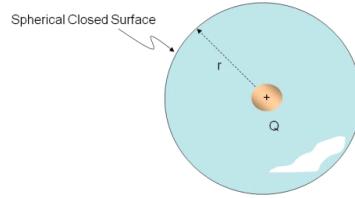
$$A = \int \int r^2 \sin \theta d\theta d\phi \quad (26.16)$$

we have to be careful not to over count area. Let's view this as first integrating around the circle of radius $r \sin \theta$ over the variable ϕ , then an integration of all these circles as θ changes from 0 to π

$$\begin{aligned} A &= \int_0^\pi \int_0^{2\pi} r^2 \sin \theta d\phi d\theta \\ &= r^2 \int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\phi \\ &= 2\pi r^2 \int_0^\pi \sin \theta d\theta \\ &= 4\pi r^2 \end{aligned} \quad (26.17)$$

as we expect.

We are now ready to do a simple problem.



Let's calculate the flux through a spherical surface if there is a point charge at the center of the sphere. The field of the point charge is

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{Q_E}{r^2} \hat{r}$$

then the flux through the surface is

$$\begin{aligned}\Phi_E &= \oint \tilde{E} \cdot d\tilde{A} \\ &= \oint \frac{1}{4\pi\epsilon_0} \frac{Q_E}{r^2} \hat{r} \cdot d\tilde{A}\end{aligned}$$

but \hat{r} is always in the same direction as $d\tilde{A}$ for this case, so

$$\hat{r} \cdot d\tilde{A} = (1)dA \cos(0) = dA$$

which gives us just

$$\begin{aligned}\Phi_E &= \frac{Q_E}{4\pi\epsilon_0} \oint \frac{1}{r^2} dA \\ &= \frac{Q_E}{4\pi\epsilon_0} \oint \frac{1}{r^2} r^2 \sin\theta d\theta d\phi \\ &= \frac{Q_E}{4\pi\epsilon_0} \int_0^\pi \left(\int_0^{2\pi} d\phi \right) \sin\theta d\theta \\ &= \frac{Q_E}{4\pi\epsilon_0} 4\pi \\ &= \frac{Q_E}{\epsilon_0}\end{aligned}$$

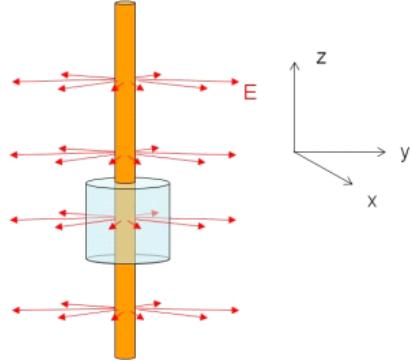
Some comments are in order. Our surfaces that we are using to calculate flux might be a real object. You might calculate the electric flux leaving a microwave oven, or a computer case to make sure you are in keeping emissions within FCC rules. But more likely the surface is purely imaginary—just something we make up.

Symmetry is going to be very important in doing problems with flux. So we will often make up very symmetrical surfaces to help us with our problems. In today's problem, the fact that \hat{r} and $d\tilde{A}$ were in the same direction made the integral *much* easier.

Until next lecture, it may not seem beneficial to invent some strange symmetrical surface and then to calculate the flux through that surface. But it is, and it will have the effect of turning a long, difficult integral into a simple one, when we can pull it off.

Flux example: a long straight wire

Let's take another example. A long straight wire.



We remember that the field from a long straight wire is approximately

$$E = \frac{1}{4\pi\epsilon_0} \frac{2|\lambda|}{r}$$

The symmetry of the field suggests an imaginary surface for measuring the flux. A cylinder matches the geometry well. Let's find the flux through an imaginary cylinder that is L tall and has a radius r and is concentric with the line of charge. Note that we are totally making up the cylindrical surface. There is not really any surface there at all.

The flux will be

$$\Phi_E = \oint \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}}$$

We can view this as three separate integrals

$$\Phi_E = \oint_{top} \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}} + \oint_{side} \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}} + \oint_{bottom} \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}}$$

since our cylinder has end caps (the top and bottom) and a curved side.

Let's consider the end caps first. For both the top and the bottom ends, $\tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}} = 0$ everywhere. No field goes thorough the ends. So there is no flux through the ends of the cylinder.

There is flux through the side of the cylinder. Note that the field is perpendicular to the side surface everywhere. So $\tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}} = EdA$. We can write our flux as

$$\begin{aligned}\Phi_E &= \oint_{side} EdA \\ &= \int \frac{1}{4\pi\epsilon_0} \frac{2|\lambda|}{r} dA\end{aligned}$$

Integrated over the side surface. But we will need an element of surface area dA for a

Question 223.26.6

cylinder side. Cylindrical coordinates seem logical so let's try

$$dA = rd\theta dz$$

then

$$\begin{aligned}\Phi_E &= \oint \oint \frac{1}{4\pi\epsilon_0} \frac{2|\lambda|}{r} rd\theta dz \\ &= \frac{2|\lambda|}{4\pi\epsilon_0} \int_0^L \int_0^{2\pi} d\theta dz \\ &= \frac{|\lambda|}{2\pi\epsilon_0} (2\pi L) \\ &= \frac{|\lambda|}{\epsilon_0} L\end{aligned}$$

So far we have, indeed, made integrals that look hard but are really easy to do. But note that this would be *much* harder if the wire were not at the center of the cylinder, or if in the previous example the charge had been off to one side of the sphere.

We would still like to remove such difficulties if we can. And often we can by choosing our imaginary surface so that the symmetry is there. But sometimes that is harder, or worse yet, we don't know exactly where the charges are in a complicated configuration of charge. We will take this on next lecture when we study a technique for finding the electric field invented by Gauss.

Basic Equations

The electric flux is defined as

$$\Phi_E = \vec{E} \cdot \vec{A} = EA \cos \theta$$

where the area vector is given by

$$\vec{A} = \hat{n}A$$

and for a curved area, we integrate

$$\Phi_E = \oint \tilde{E} \cdot d\tilde{A}$$

27 Gauss' Law and its Applications

Fundamental Concepts

- Gauss' Law tells us that the flux through a closed surface is equal to the charge inside the surface divided by ϵ_o :

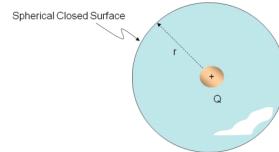
$$\Phi = \frac{Q_{in}}{\epsilon_o}$$

- Gauss' Law combined with our basic flux equation

$$\Phi_E = \oint \vec{E} \cdot d\vec{A} = \frac{Q_E}{\epsilon_o}$$

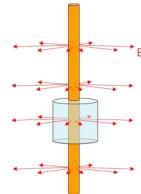
Gauss' Law

Last lecture we did two problems. We found the flux from a point charge through a spherical surface to be



$$\Phi_{sphere,point} = \frac{Q_E}{\epsilon_o}$$

and the flux from a line of charge through a cylinder to be



$$\Phi_{cylinder,line} = \frac{|\lambda|}{\epsilon_o} L$$

Let's rewrite the last one using

$$\lambda = \frac{Q}{L}$$

then

$$\begin{aligned}\Phi_{cylinder,line} &= \frac{|Q_E/L|}{\epsilon_0} L \\ &= \frac{|Q_E|}{\epsilon_0}\end{aligned}$$

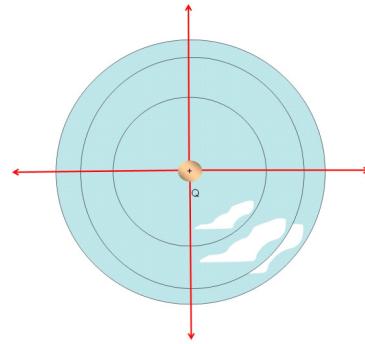
which is just what we got for the point charge and sphere! That is amazing! Think about how much work it was to find each flux, and in the end we got the same result. Wouldn't it be great if the flux through every closed surface was this simple? Then we would not have to integrate at all!

To see if we can do this, first let's think of our answer.

$$\Phi_{sphere,point} = \frac{Q_E}{\epsilon_0}$$

Question 223.27.1

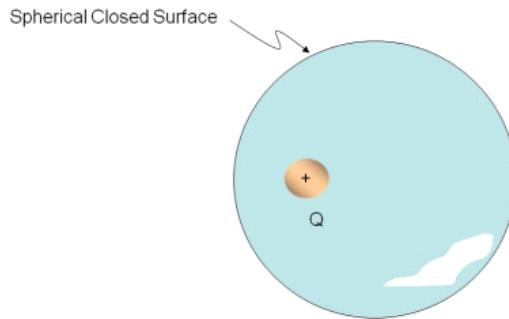
It does not depend on the radius of the spherical surface. So any spherical surface centered on the charge will do! This makes sense. No matter how big the sphere, all the field lines must leave it. Since flux gives the amount of field that penetrates an area, for our charge at the center of a sphere we see that all of the field penetrates the spherical surface no matter the size of the sphere. So the flux is the same no matter r .²¹



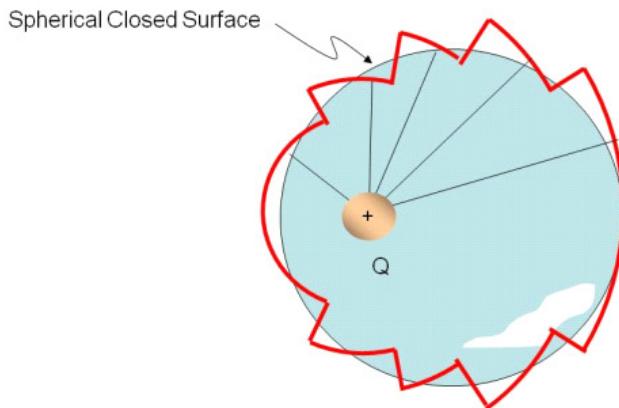
The key to making our last lecture problems easy was that the field was always perpendicular to the surface so $\vec{E} \cdot d\vec{A} = EdA$ was easy to find.

Using geometry we can arrange to make nearly all of our flux problems like this. To demonstrate, let's take the case of a point charge that is off center in a spherical surface.

²¹ If this still seems strange, remember that the area of a sphere is $4\pi r^2$ and that the field of a point charge is $\frac{1}{4\pi\epsilon_0} \frac{Q}{r^2}$. The flux is like the product of these two quantities. The r^2 terms must cancel. So the fact that the flux is the same for any sphere is due to the r^2 dependence of the field.



Remember, we made up this surface. So we can place the surface anywhere we like. And this time we would like the charge to be off center. We will call these made up surfaces *Gaussian surfaces* after the mathematician that thought up this method of avoiding integrals. Having the charge off center would make for a difficult integration because \vec{E} and $d\vec{A}$ have different directions as we go around the sphere. But let's consider, would there be less flux through the surface than there was when the charge was centered in the sphere? Every field line that is generated will still leave the surface. Flux gives us the amount of field that penetrates the surface.²² Since flux is the amount of field penetrating our surface, it seems that the flux should be exactly the same as when the charge was in the center of the sphere. To prove this, let's take our surface and approximate it using area segments. But let's have the area segments be either along a radius of a sphere centered on the charge, or along the surface of a sphere centered on the charge.



No flux goes through the radial pieces. And the rest of the pieces are all parts of spheres centered on the charge. But for the spherical segments, the field will be perpendicular

²² Think of water flow rate again. We could place the end of a garden hose in a wire mesh container. The water would flow out the hose end and through the wire mesh sides of the container. The flow rate tells us how much water passes through the container surface. The flow rate does not depend on the shape of the container. The hose end is like a charge. The hose is the source of water, the charge is the source of electric field.

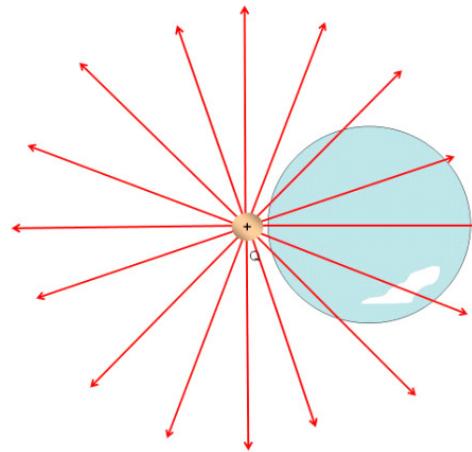
to the segment no matter what sphere the segment is a part of, because we chose only spheres that were concentric with the charge. The r we have for the little spherical pieces does not matter, so on all of these surfaces $\vec{E} \cdot d\vec{A} = EdA$. Then the integration for these pieces will be easy.

Of course this surface made of little segments from other spheres is a poor approximation to the shape of the offset sphere. But we can make our small segments smaller and smaller. In the limit that they are infinitely small, our shape becomes the offset sphere. That means that once again our flux is

$$\Phi = \frac{Q_E}{\epsilon_o}$$

This is fantastic! We don't have to do the integration at all. We just count up the charge inside our surface and divide by ϵ_o .

What happens if the charge is on the outside of the surface?

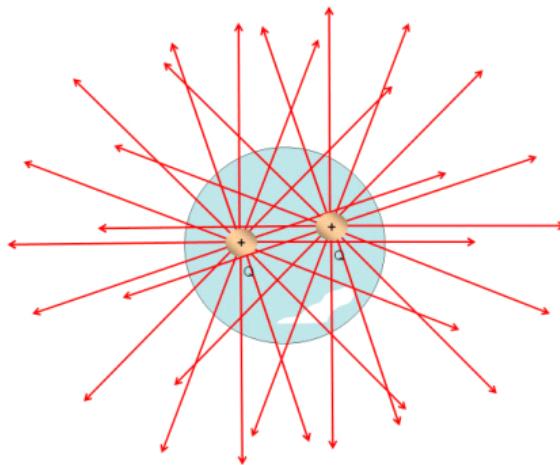


Every field line that enters goes back out. We encountered this last time. The flux going in is negative, the flux going out is positive, and they must be the same because every line leaves that enters. So the net flux must be zero. This means we should still write our flux as

$$\Phi = \frac{Q_{inside}}{\epsilon_o}$$

because outside charges won't contribute to the flux. So in a way, our expression works for charges outside our closed surface.

We know that fields superimpose, that is, they add up, so we would expect that if we have two charges inside a surface,



we would add up their contributions to the total flux

$$\Phi_{total} = \Phi_1 + \Phi_2$$

which means that Q_{inside} is the sum of all the charges inside. We recognize that if some charges are negative, they will cancel equal amounts of charge that are positive.

This leaves us with a fantastic time savings law

The electric flux Φ through any closed surface is equal to the net charge inside the surface multiplied by $4\pi k_e$. The closed surface is often called a *Gaussian Surface*.

$$\Phi_E = \oint \vec{E} \cdot d\vec{A} = \frac{Q_{inside}}{\epsilon_0} \quad (27.1)$$

This was first expressed by Gauss, and therefore this expression is called Gauss' law.

Examples of Gauss' Law

Question 223.27.2

But why do we get so excited about flux? The reason is that we can use the idea of flux combined with Gauss' law gives us an easy way to calculate the electric field from a distribution of charge if we can find a suitable symmetric surface! If we can find the field, we can find forces, and we can predict motion.

Let's show how to do this by working some examples.

Charged Spherical Shell

First let's take a charged spherical shell and find the field inside. We need to be able to guess the shape of the field. We use symmetry. We can guess that the field will be radial

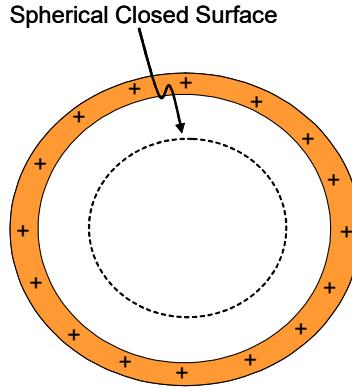


Figure 27.17.

both inside and outside of the shell. If it were not so, then our symmetry tests would fail.

The shell has a total charge of $+Q$. If we place a spherical surface inside the shell, then we can use Gauss's law.

$$\Phi = \frac{Q_{inside}}{\epsilon_0}$$

We can tell from the symmetry of the situation that \vec{E} is everywhere colinear with (but in the opposite direction as) $d\vec{A}$ so

$$\Phi = \oint \vec{E} \cdot d\vec{A} = - \oint E dA$$

because the field is everywhere perpendicular to the surface. We can even make a guess that the field must be constant on this surface, because all along the spherical Gaussian surface there is extreme symmetry. No change in reflection, or rotation etc. will change the shape of the charge, so around the spherical surface the field must have the same value. Then

$$\Phi = -E \oint dA = -EA$$

Equating our flux equations gives

$$-EA = \frac{Q_{inside}}{\epsilon_0}$$

or

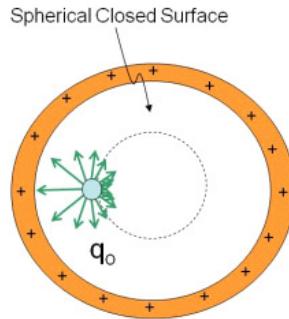
$$E = -\frac{Q_{inside}}{A\epsilon_0}$$

but what is Q_{inside} ? It is zero! so

$$E = -\frac{0}{A\epsilon_0} = 0$$

There is no net field inside!

This may seem surprising, but think of placing a test charge, q_o , inside the sphere. The next figure shows the forces acting on such a test charge. The force is stronger between the charge and the near surface, but there is more of the surface tugging the other way.



The forces just balance. Since

$$F = qE$$

Question 223.27.3

if the net force is zero, then the field must be zero too.

Is there a field outside of the spherical shell? It is still true that

$$\Phi = \oint \vec{E} \cdot d\vec{A} = \oint E dA$$

but this time we have a positive sign on the last integral because \vec{E} and $d\vec{A}$ are in the same direction. Then

$$EA = +\frac{Q_{inside}}{\epsilon_0}$$

We now choose our surface around the entire shell. All of our analysis is the same as in

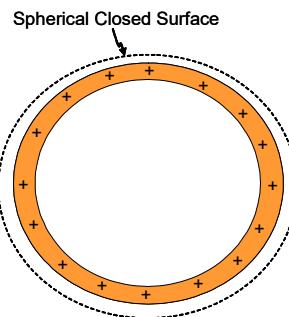


Figure 27.18.

our last problem, except now Q_{inside} is not zero

$$E = \frac{Q_{inside}}{A\epsilon_0}$$

The area is the area of our imaginary sphere

$$E = \frac{Q_{inside}}{(4\pi r^2) \epsilon_0}$$

and since $Q_{inside} = +Q$, then

$$E = \frac{+Q}{4\pi\epsilon_0 r^2}$$

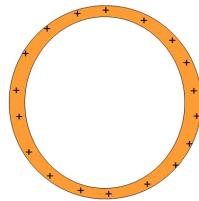
and we have found the field.

Note that this field looks very like a point charge at the center of the spherical shell (at the center of charge), but by now that is not much of a surprise!

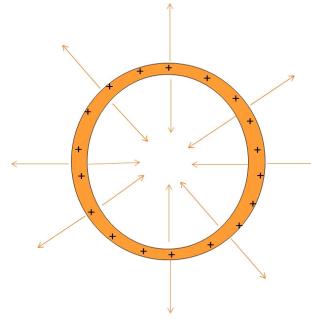
Strategy for Gauss' law problems

Let's review what we have done before we go on to our last example. For each Gauss' law problem, we

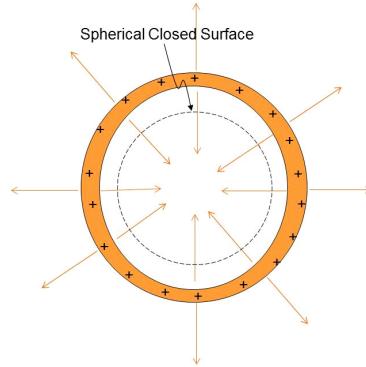
1. draw the charge distribution



2. Draw the field lines using symmetry



3. Choose (make up, invent) a closed surface that makes $\vec{E} \cdot d\vec{A}$ either just $E dA$ or 0.



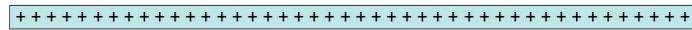
4. Find Q_{in} .

5. Solve $\oint EdA = \frac{Q_{inside}}{\epsilon_0}$ for the non, zero parts

The integral should be trivial now due to our use of symmetry.

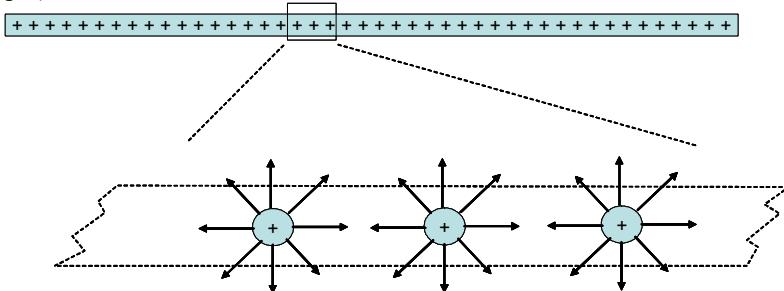
An infinite sheet of charge.

Spherical cases were easy. Let's try a harder one. Let's try our infinite sheet of charge. It is a little hard to draw. So we will draw it looking at it from the side from within the sheet of charge (somewhere in it's middle, if an infinite sheet can have a middle).

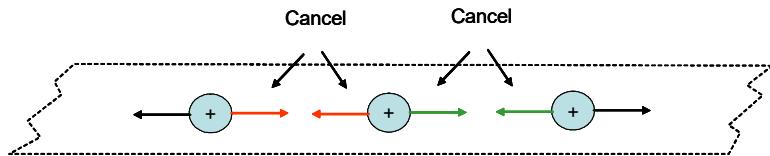


This completes step 1).

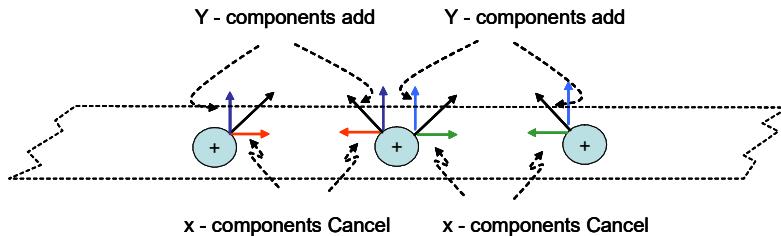
For step 2), let's think about what the electric field will look like.



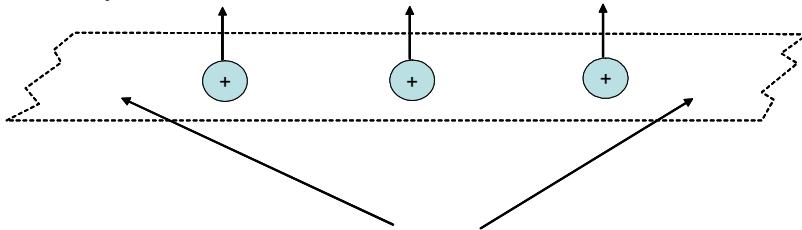
In the figure above I have blown up the view on three charge carriers and drawn some field lines. Notice that in the x -direction the fields will cancel.



The y -components add

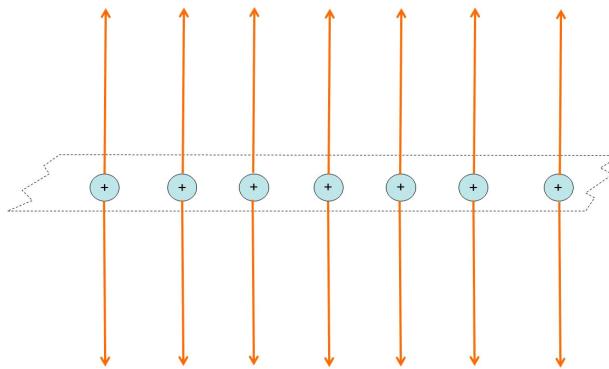


So we have only a field in the y direction



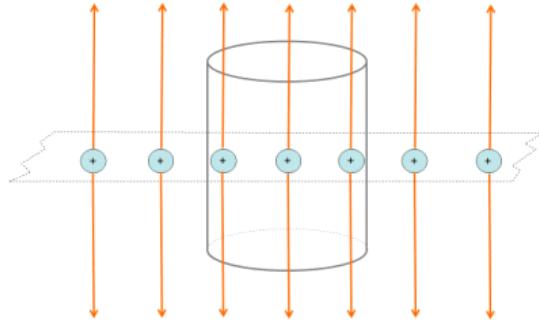
Remember that this only works if we have the rest of the sheet to cancel the components on the end charges shown

Now if we had edges of our sheet of charge, not all the x -components would cancel and the problem would be harder, but we won't do that problem now. Also note that there is a field in the $-y$ -direction, I only drew some of the field lines in the figures.



This is step 2).

Now we need to choose an imaginary surface over which to integrate $\oint \vec{E} \cdot d\vec{A}$. We want $\vec{E} \cdot d\vec{A} = EdA$ or $\vec{E} \cdot d\vec{A} = 0$ over all parts of the surface. I suggest a cylinder.



Note that along the top of the cylinder, $E \parallel A$ so $\vec{E} \cdot d\vec{A} = EdA \cos \theta = EdA$. Along the side of the cylinder $E \perp A$ so $\vec{E} \cdot d\vec{A} = EdA \cos \theta = 0$. We have a surface that works! This completes step 3).

Now we need to solve the integral. The flux is just

$$\begin{aligned}\Phi &= \oint \vec{E} \cdot d\vec{A} \\ \Phi &= \oint_{\text{side}} \vec{E} \cdot d\vec{A} + \oint_{\text{ends}} \vec{E} \cdot d\vec{A} \\ &= 0 + \oint_{\text{ends}} EdA = 2EA\end{aligned}$$

where the factor of 2 comes because we have two caps and field in the $+y$ and $-y$ directions and where A is the area of one end cap. If we know that the sheet of charge has a surface charge density of η , then we can write the charge enclosed by the cylinder as

$$Q_{\text{inside}} = \eta A$$

so

$$\Phi_E = \frac{\eta A}{\epsilon_0}$$

by Gauss' law. Equating the two expressions for the flux gives

$$2EA = \frac{\eta A}{\epsilon_0}$$

or

$$E = \frac{\eta}{2\epsilon_0} \quad (27.2)$$

which is what we found before for an infinite sheet of charge, but this way was *much* easier. If we can find a suitable surface, Gauss' law is very powerful!

Gauss's law strategy

In each of our problems today, we found the electric field without a nasty integration. Usually we want the electric field at a specific point. To make Gauss' law work we need

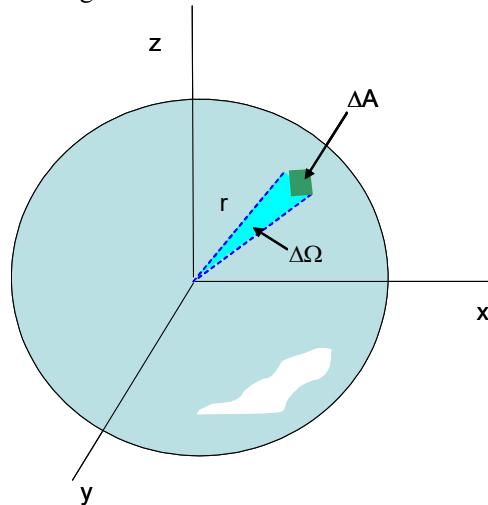
to do the following for each problem:

1. Draw the charge distribution
2. Draw the field using symmetry
3. Invent a Gaussian surface that takes advantage of the field symmetry and that includes our point where we want the field. We will want $\vec{E} \cdot d\vec{A} = EdA$ or $\vec{E} \cdot d\vec{A} = 0$ for each part of the surface we invent.
4. Find the flux by finding the enclosed charge, Q_{in}

Question 223.27.4
 5. use $\oint \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\epsilon_0}$ integrating over our carefully invented surface to find the field. If our surface that we imagined was good, then $\oint \vec{E} \cdot d\vec{A}$ will be very easy.

Derivation of Gauss' Law

A formal derivation of Gauss' Law is instructive, and it gives us the opportunity to introduce the idea of solid angle.



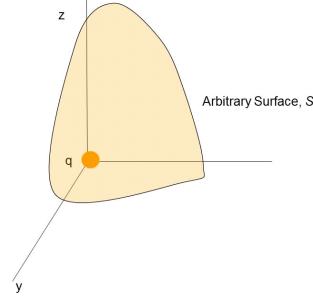
$$\Delta\Omega = \frac{\Delta A}{r^2} \quad (27.3)$$

This is like a two dimensional angle. And just like an angle, it really does not have dimensions. Note that ΔA is a length squared, but so is r^2 . The (dimensionless) unit for solid angle is the *steradian*. We can see that for a sphere we would have a total solid

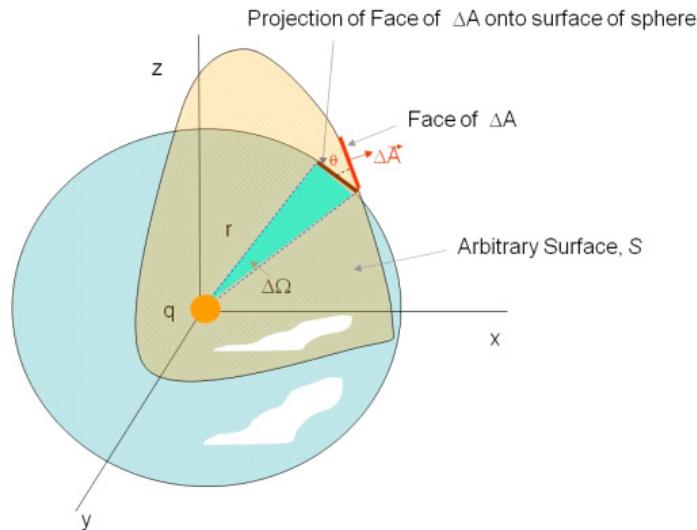
angle of

$$\Omega_{sphere} = \frac{4\pi r^2}{r^2} = 4\pi \text{ sr} \quad (27.4)$$

Now let's see why this is useful. Consider a point charge in an arbitrary closed surface.



If we look at a particular element of surface ΔA we can find the flux through that surface element. We can use our idea of solid angle to do this



$$\Delta\Phi_E = \tilde{\mathbf{E}} \cdot \Delta\tilde{\mathbf{A}}$$

Since the field lines are symmetric about q and the surface is arbitrary, the element $\Delta\tilde{\mathbf{A}}$ will be at some angle θ from the field direction so

$$\tilde{\mathbf{E}} \cdot \Delta\tilde{\mathbf{A}} = E\Delta A \cos\theta$$

this is no surprise. But now notice that the projection of ΔA puts it onto a spherical

surface of just about the same distance from q . The projected area is

$$\Delta A_P = \Delta A \cos \theta$$

At this point we should remember that we know the field due to a point charge

$$E = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$$

so our flux through the area element is

$$\begin{aligned}\Delta\Phi_E &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \Delta A \cos \theta \\ &= \frac{q}{4\pi\epsilon_0} \frac{\Delta A \cos \theta}{r^2}\end{aligned}$$

but

$$\frac{\Delta A \cos \theta}{r^2} = \Delta\Omega$$

is the solid angle subtended by the projected area. Then

$$\Delta\Phi_E = \frac{q}{4\pi\epsilon_0} \Delta\Omega$$

The total flux though the oddly shaped closed surface is then

$$\Phi_E = \frac{q}{4\pi\epsilon_0} \oint d\Omega$$

where we integrate over the entire arbitrary surface, S .

$$\Phi_E = \frac{q}{4\pi\epsilon_0} \oint_S d\Omega$$

but by definition

$$\oint_S d\Omega = 4\pi \text{ sr}$$

so

$$\begin{aligned}\Phi_E &= \frac{q}{4\pi\epsilon_0} \oint_S d\Omega \\ &= \frac{q}{4\pi\epsilon_0} 4\pi \text{ sr} \\ &= \frac{q}{\epsilon_0}\end{aligned}$$

which is just Gauss' law.

So far we have used mostly charged insulators to find fields. But we know we will be interested in conductors and their fields in building electronics. We will take up the study of charged conductors and their fields next.

Basic Equations

Gauss' law

$$\Phi = \frac{Q_{inside}}{\epsilon_0}$$

Gauss' law combined with our equation for flu

$$\Phi = \oint \vec{E} \cdot d\vec{A} = \frac{Q_{inside}}{\epsilon_0}$$

28 Conductors in Equilibrium, Electric Potentials

Fundamental Concepts

- Conductors in Equilibrium
- Electric Potential Energy

Conductors in Equilibrium

Conductors have some special properties because they have movable charge. Here they are

1. Any excess static charge (charge added to an uncharged conductor) will stay on the surface of the conductor.
2. The electric field is zero everywhere *inside* a conductor.
3. The electric field just outside a charged conductor is perpendicular to the conductor surface.
4. Charge tends to accumulate at sharp points where the radius of curvature of the surface is smallest.

It is our job to convince ourselves that these are true. Lets take these one at a time.

In Equilibrium, excess charge is on the Surface

Question 223.28.1

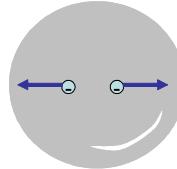
Let's think about what we know about conductors. Most good conductors are metals. The reason they are good conductors is that the outer electrons in metals are in open valence bands where there are many energy states available to the electrons. These electrons are free to travel around. This means that if we place a charge near a metal object, the free charges will experience an acceleration. Of course, the charge does not

fly out of the conductor. It will have to stop when it reaches the end of the metal object. Suppose we go back to our experiment from the first lecture. We took a charged rod, and placed it near an uncharged conductor.



The free electrons moved. We ended up with a bunch of electrons all on the right hand side. They all repel each other. So at some point the force between a free electron and the charged rod, and the force between a free electrons and the rest of the free electrons will balance. At that point, there is zero net force (think of Newton's second law). The free electrons stop moving. We have a word from PH121 or Statics for when all the forces balance. We say the charges are in *equilibrium*.

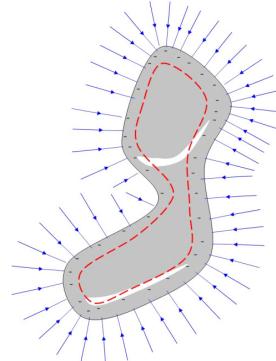
Now suppose we have a conductor just on it's own and suppose we add charge to it. Where would the extra charge go? We have considered this before. In the picture below, I have a spherical conductor with two extra negative charges shown. The pair of charges will repel each other. Now because of the r^2 in our electric force equation, the closer the extra charges are, the stronger the repulsive force. The result is that they will try to go as far from each other as possible. So the extra charge on a spherical conductor will all end up on the surface.



The Electric Field is Zero Inside a Conductor

Question 223.28.2

We can use Gauss' law to find the field in a conductor. We know that the extra charge will all be on the surface if there is no electric current.

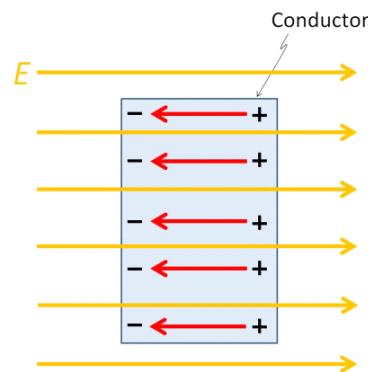


We can then draw a Gaussian surface, to match the symmetry of the conductor. What is the charge inside the Gaussian surface? It is net zero, since the reaming charge is all bound up in atoms and balances out. Since there is no net charge, there is no net flux. If there is no flux, there is no net field inside a conductor that is in static equilibrium.

Note that if we connected this conductor to both ends of a battery, we would have a field in the conductor generated by the battery and the charge flow it creates, so we must remember that static equilibrium is a special case.

If we don't connect the conductor to the ground or a battery, we can say: *The net electric field is zero everywhere inside the conducting material.*

Consider if this were not true! If there were an electric field inside the conductor, the free charge there would accelerate and there would be a flow of charge. If there were a movement of charge, the conductor would not be in equilibrium. Suppose we place a brick of conductor in a field. We expect that the charges will be accelerated. Negative charges will move opposite the field direction. We end up with the situation shown in the next figure.



Since the negative charges moved, the other side has a net positive charge. This

separation of the charges creates a new field in the opposite direction of the original field. In equilibrium, just enough charge is moved to create a field that cancels the original field.

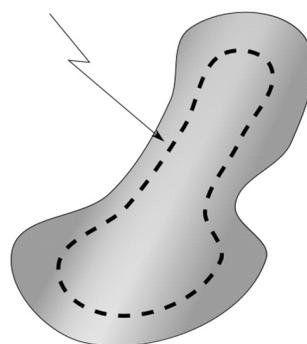
Return to charge being on the surface

Question 223.28.3

Suppose we have a conductor in equilibrium. We can now ask, what does it mean that the charge is “on the surface?” Is there a small distance within the metal where we would find extra charge? or is it all right at the edge of the metal?

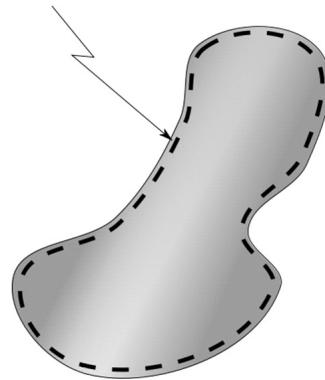
Let's look at this again now that we know Gauss' law. Let's envision a conducting object with a matching Gaussian surface.

Closed Gaussian Surface



We know the field inside the conductor is zero. So no field lines can leave or enter the Gaussian surface. So no charge can be inside or we would have a net flux, and, therefore, a field. We can move the Gaussian surface from the center of the conductor and grow it until it is just barely smaller than the surface of the conductor, and there still must be no field, so no charge inside.

Closed Gaussian Surface

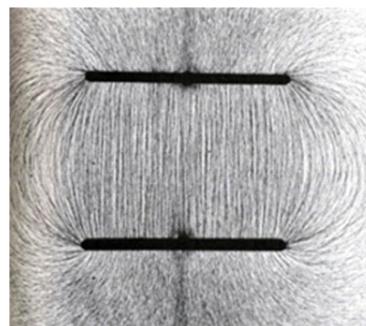


We can make this Gaussian surface as close to the actual surface as we like, and still there must be no field inside. Thus all the excess charge must be on the surface. It is not distributed at any depth in the material.²³

Field lines leave normal to the surface

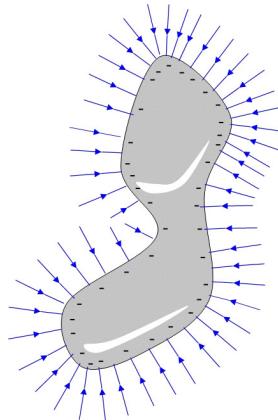
Question 223.28.4

In the following picture, we can see that the field lines seem to leave the surface of these charged conductors at right angles (remember that sometimes we call this *normal* to the surface).

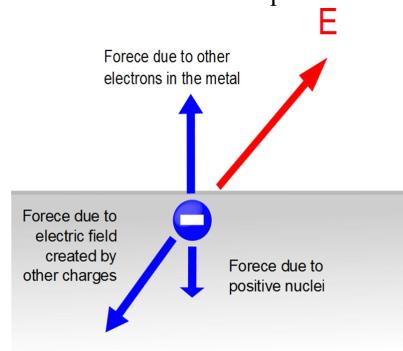


We have charges all along the surface, and neighboring charges cancel all but the normal components of the field, so the field lines go straight out. Notice that farther from the conductor the field lines may bend, but they start out leaving the surface perpendicular to the surface. Let's draw a conducting object.

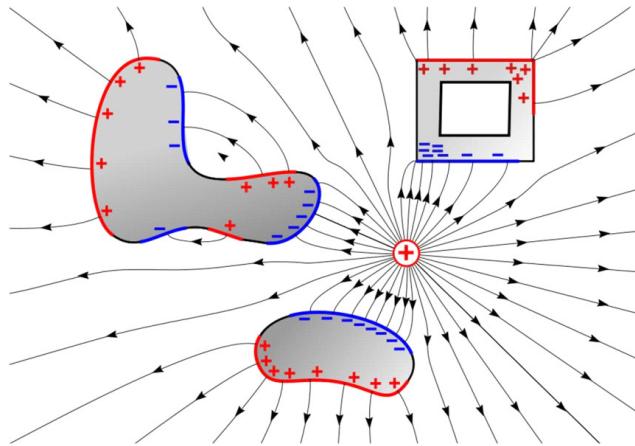
²³ For our chemists, our quantum picture will modify this reasoning a little, since we will view electrons as waves that extend out into space a bit.



Consider what would happen if it were not true that the field lines left perpendicular to a conductor surface when the conductor was in equilibrium.

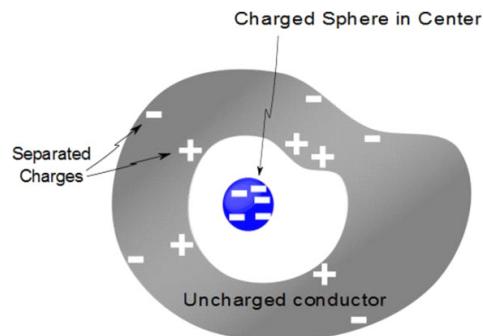


There would be a horizontal component of the field in such a case. The component of the field along the surface would cause the charge to move. In the figure there would be a net force to the left. This force would rearrange the charge until there was no force. But since $F_x = qE_x$, then when F_x is zero, so is E_x . Suppose we place a conductor in an external field. We would see that the charges within the conductor will rearrange themselves until the field lines will leave perpendicular to the surface of the conductors.



Notice the square box in the last figure. There is an opening inside the conductor, but there is no net field inside. The conductor charges rearrange themselves so that the external field is canceled out. This is part of what is known as a *Faraday cage* which allows us to cancel out an external electric field. This is used to protect electronic devices that must operate in strong electric fields. To complete the effect, we will also need to show that magnetic fields are canceled by such a conducting box.

We should also consider what happens when we place a charge in a conductive container. Does this charge get screened off? That is, would the conductive container prevent us from telling if there was a charge inside?

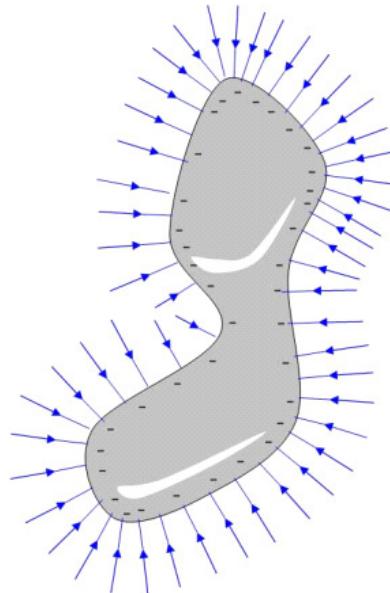


In this case, the answer is no. The charges in the conductor will move because of the charge contained inside the conducting container. The negative charge will move as shown, and it will move to the outside of the container surface. This leaves positive charges behind on the inner surface. We know that there will be no field inside the

conductor material, itself. But think of placing a Gaussian surface around all of the container and charge. There will be a net charge inside the Gaussian surface, so there will be a field. The inner surface charge does cancel the charge from the charged sphere. But the negative charge on the conductor surface creates a new field.

Charge tends to accumulate at sharp points

Let's go back to our charged conductor. Notice that the field lines bunch up at the corners! Where the field lines are closer together, there must be more charge and the field strength must be higher.



Now that we have an idea of how charge and conductors act in equilibrium, we would like to motivate charge to move. To see how this happens, let's review energy.

Electrical Work and Energy

Question 223.28.5

Put this on the far board

We remember studying energy back in PH121 or Statics and Dynamics. Remember the Work-Energy theorem?

$$W_{nc} = \Delta K + \Delta U \quad (28.1)$$

We started with gravitational potential energy, and, as we found conservative forces, we defined new potential energies to describe the work done by those forces. For example,

we added spring potential energy

$$W_{nc} = \Delta K + \Delta U_g + \Delta U_s \quad (28.2)$$

I bet you can guess what we will do with our electrical or Coulomb force!

$$W_{nc} = \Delta K + \Delta U_g + \Delta U_s + \Delta U_C \quad (28.3)$$

When we do this, we mean that the work done by the Coulomb force (W_C) is the negative of the electrical potential energy change

$$W_C = -\Delta U_C \quad (28.4)$$

and we are saying that the Coulomb force is conservative. But is the Coulomb force conservative? Remember that the equation for the force due to gravity and the equation for the Coulomb force are very alike. So we might guess that the Coulomb force is conservative like gravity—and we would be right!

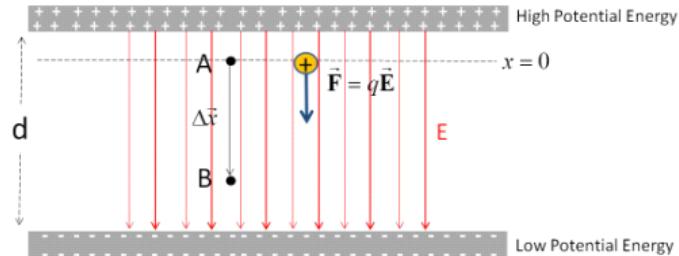
Energy of a Charge in a uniform field

Question 223.28.6

Question 223.28.7

Question 223.28.8

Let's use our Coulomb force to calculate work. I would like a simple example, so let's assume we have a uniform electric field. We know that we can almost really make a uniform electric field by building a large capacitor.



We draw some field lines (from the + charges to the - charges). The field lines will be mostly straight lines in between the plates. Of course, outside the plates, they will not be at all straight, but we will ignore this because we want to calculate work just in the uniform part of the field.

I want to place a charge, q , in this uniform field. The charge will accelerate. Work will be done. I want to find out how much work is done on the charge.

From our PH121 or Dynamics experience, we know that

$$\begin{aligned} W &= \int \vec{F} \cdot d\vec{x} \\ &= F\Delta x \cos \theta \end{aligned} \quad (28.5)$$

for constant forces. Because we have a constant field, we will have a constant force.

I will choose the x direction to be vertical and $x = 0$ to be near the positive plate. Then we can write the force due to the electric field as

$$\begin{aligned} W &= F\Delta x \cos \theta \\ &= (q_m E) \Delta x \cos(0^\circ) \\ &= q_m E \Delta x \end{aligned}$$

Put this on the far board

If there are no non-conservative forces, and we ignore gravity, then we can say

$$\begin{aligned} W_{nc} &= \Delta K + \Delta U_g + \Delta U_s + \Delta U_C \\ 0 &= \Delta K + 0 + \Delta P E_C \\ 0 &= \Delta K + 0 - q_m E \Delta x \end{aligned}$$

so

$$\Delta K = q_m E \Delta x \quad (28.6)$$

This is very interesting! This means that for this simple geometry I could ask you questions like, “after the charge travels Δx , how fast is it going?”

Electric and Gravitational potential energy compared

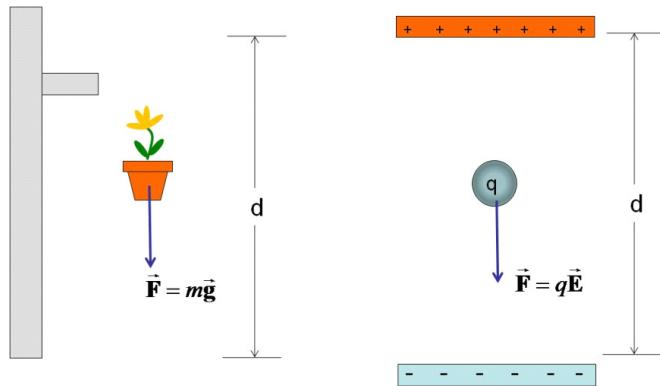
We have found that the potential energy for the Coulomb force is given by

$$\Delta U_C = -q_m E \Delta x$$

for a *uniform* electric field (it will change for non-uniform fields). Let’s compare this to the gravitational potential energy

$$\Delta U_g = -mgh$$

Let’s set up a situation where the electric field and gravitational field are almost uniform and we have a positively charged particle with charge q and mass m . The height, h , we will call d to match our gravitational and electrical cases.



The gravitational potential difference is

$$\Delta U_g = -mgd \quad (28.7)$$

and the electrical potential difference is

$$\Delta U_C = -q_m Ed \quad (28.8)$$

These equations look a lot alike. We should expect that if we push the charge q_m "up," we will increase both potential energies. We will have to do positive work to do that ($W = -\Delta U$). This is just like doing work in a gravitational field, so we are familiar with this behavior.

There is a difference, however. We have assumed that our charge q_m was positive. Suppose it is negative? There is only one kind of mass, but we have two kinds of charge. We will have to get used to negative charges "falling up" to make the analogy continue.

This analogy helps us to understand how the electric potential energy will act, and we will continue to use it. There is a difficulty, however, in that most engineering classes only study gravitation in nearly uniform gravitational fields. But if we look at large objects (like whole planets) that are separated from other objects by some distance, then we have very non-uniform gravitational fields. Unless you are an aerospace engineer, these cases are less common. So to help us understand electric potential energy, we will study gravitational potential energy of large things first, then study the energy associated with individual charges and their very non-uniform fields. We will take this on next time.

Basic Equations

29 Electric potential Energy

Fundamental Concepts

- Gravitational potential energy of point masses and binding energy
- Electrical potential energy of point charges
- Electrical potential energy of dipoles

Point charge potential energy

As we said last lecture, we want to use gravitation as an analogy for the electric potential energy. Gravitation is more intuitive. But chances are gravitation of whole planets was not stressed in Dynamics (If you took PH121 you should be fine, and this will be a review). So let's take a few moments out of a PE101 class (introductory planetary engineering) and study non-uniform gravitational fields.

Gravitational analog

Question 223.29.1

Question 223.29.2

Long, long ago you studied the potential energy of objects in what we can now call the Earth's gravitational field.

The presentation of the idea of potential energy likely started with

$$U_g = mgy$$

where m is the mass of the object, g is the acceleration due to gravity, and y is how high the object is compared to a $y = 0$ point. If you recall, we got to pick that $y = 0$ point. It could be any height.

This all works fairly well so long as we take fairly small objects near the much larger Earth. But hopefully you also considered objects farther away from the Earth's surface, or larger objects like the moon. For these objects, mgy is not enough to describe the potential energy. The reason is that if we are far away from the center of the Earth we

will notice that the Earth's gravitational field²⁴ is not uniform. It curves and diminishes with distance. So, if an object is large, it will feel the change in the gravitational field over its (the object's) large volume.

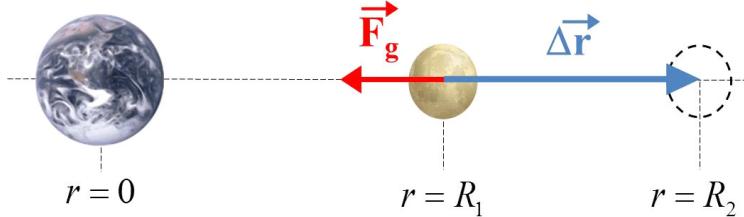
We have the tools to find the potential energy of this situation. We know that a change in potential energy is just an amount of work

$$\Delta U_g = -W_g = - \int \vec{F}_g \cdot d\vec{r}$$

The magnitude of the gravitational force is

$$F_g = G \frac{M_E m_m}{r_{Em}^2}$$

where M_E is the mass of the Earth, m_m is the mass of the mover object, and r_{Em} is the distance between the two. The constant, G , is the gravitational constant.



The field is radial, so $\vec{F}_g \cdot d\vec{r} = -Fdr$ for the configuration we have shown, and we can perform the integration. Say we move the object a distance Δr away were

$$\Delta r = R_2 - R_1$$

and Δr is large, comparable to the size of the Earth or larger. Then

$$\begin{aligned} \Delta U_g &= - \int_{R_1}^{R_2} \left(-G \frac{M_E m_m}{r^2} \right) dr \\ &= GM_E m_m \int_{R_1}^{R_2} \frac{dr}{r^2} \end{aligned}$$

²⁴ Of course, the gravitational field is really the warping of space-time. But that is a subject for another physics class.

where R is the distance from the center of the Earth to the center of our object.

$$\begin{aligned}\Delta U_g &= GM_E m_m \int_{R_1}^{R_2} \frac{dr}{r^2} \\ &= GM_E m_m \left[-\frac{1}{r} \right]_{R_1}^{R_2} \\ &= GM_E m_m \left[-\frac{1}{R_2} - \left(-\frac{1}{R_1} \right) \right] \\ &= -GM_E m_m \left[\frac{1}{R_2} - \frac{1}{R_1} \right] \\ &= -G \frac{M_E m_m}{R_2} + G \frac{M_E m_m}{R_1}\end{aligned}$$

We recall that we need to set a zero point for the potential energy. Before, when we used the approximation $m_m g y$ we could choose $y = 0$ anywhere we wanted. But now we see an obvious choice for the zero point of the potential energy. If we let $R_2 \rightarrow \infty$ and then the first term in our expression will be zero. Likewise, if we let $R_1 \rightarrow \infty$ the second term will be zero. It looks like as we get infinitely far away from the Earth, the potential energy naturally goes to zero! Mathematically this makes sense. But we will have to interpret what this choice of zero point means.

But first, let's see how much work it would take to move the moon out of orbit and move it farther away. Say, from R_1 , the present orbit radius, to $R_2 = 2R_1$, or twice the original orbit distance. Then

$$\begin{aligned}\Delta U_g &= U_2 - U_1 = -G \frac{M_E m_m}{2R_1} + G \frac{M_E m_m}{R_1} \\ &= G \frac{M_E m_m}{R_1} \left(-\frac{1}{2} + 1 \right) \\ &= \left(\frac{1}{2} \right) G \frac{M_E m_m}{R_1}\end{aligned}$$

The change is positive. We gained potential energy as we went farther from the Earth's surface. That makes sense! That is analogous to increasing y in mgy . The potential energy also gets larger if the mass of our object (like the moon or a satellite) gets larger. Again that makes sense because in our more familiar approximation the potential energy increases with mass. So this new form for our equation for potential energy seems to work.

But what does it mean that the potential energy is zero infinitely far away? Recall that a change in potential energy is an amount of work

$$W = -\Delta U$$

Usually we will consider the potential energy to be the amount of work it takes to bring the test mass m_m from infinitely far away (our zero point!) to the location where we want it. It is how much energy is stored by having the object in that position. Like how much energy is stored by putting a mass high on a shelf. For example we could bring the moon in from infinitely far away. Then

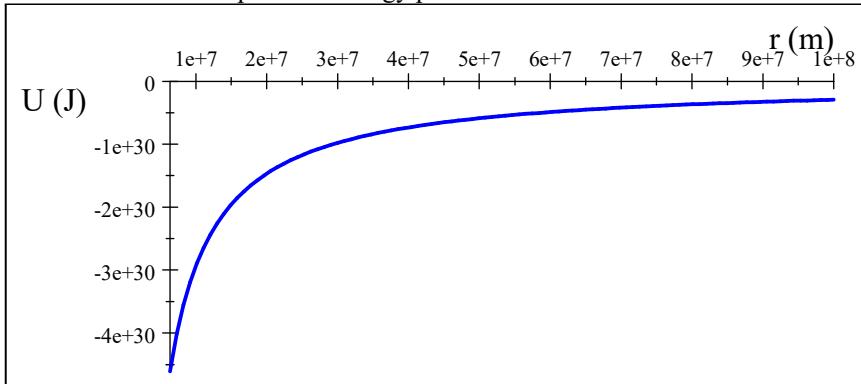
$$\Delta U_g = U_2 - U_1 = -G \frac{M_E m_m}{R_2} + G \frac{M_E m_m}{\infty}$$

$$U_2 = -G \frac{M_E m_m}{R_2}$$

This is how much potential energy the moon has as it orbits the Earth because it is high, above the Earth. But notice, this is a negative number! What can it mean to have a negative potential energy?

Question 223.29.3

We use this convention to indicate that the test mass, m_m is bound to the Earth. It would take an input of energy to get the moon free from the gravitational pull of the Earth. Here is the Moon potential energy plotted as a function of distance.



We can see that you have to go an infinite distance to overcome the Earth's gravity completely. That makes sense from our force equation. The force only goes to zero infinitely far away. When we finally get infinitely far away, there will be no potential energy due to the gravitational force because the gravitational force will be zero.

Of course, there are more than just two objects (Earth and Moon) in the universe, so as we get farther away from the Earth, the gravitational pull of, say, a galaxy, might dominate. So we might not notice the weak pull of the Earth as we encounter other objects.

We should show that this form for the potential energy due to gravity becomes the more familiar mgh if our distances are small compared to the Earth's radius.

Let our distance from the center of the Earth be $R_2 = R_E + y$ where R_E is the radius

of the Earth and $y \ll R_E$. Then

$$\begin{aligned} U &= -G \frac{M_E m_m}{R_2} \\ &= -G \frac{M_E m_m}{R_E + y} \end{aligned}$$

We can rewrite this as

$$\begin{aligned} U &= -G \frac{M_E m_m}{R_E \left(1 + \frac{y}{R_E}\right)} \\ &= -G \frac{M_E m_m}{R_E} \left(1 + \frac{y}{R_E}\right)^{-1} \end{aligned}$$

Since y is small y/R_E is very small and we can approximate the term in parenthesis using the binomial expansion

$$(1 \pm x)^n \approx 1 \mp nx \quad \text{if } x \ll 1$$

then we have

$$\left(1 + \frac{y}{R_E}\right)^{-1} \approx 1 - (-1) \frac{y}{R_E} \quad \text{if } \frac{y}{R_E} \ll 1$$

and our potential energy is

$$U = -G \frac{M_E m_m}{R_E} \left(1 + \frac{y}{R_E}\right)$$

then

$$\begin{aligned} U &= -G \frac{M_E m_m}{R_E} + G \frac{M_E m_m y}{R_E^2} \\ &= U_o + m_m \left(G \frac{M_E}{R_E^2}\right) y \end{aligned}$$

If we realize that U_o is the potential energy of the object at the surface of the Earth, then the change in potential energy as we lift the object from the surface to a height y is

$$\begin{aligned} \Delta U &= \left(U_o + m_m \left(G \frac{M_E}{R_E^2}\right) y - \left(U_o + m_m \left(G \frac{M_E}{R_E^2}\right) (0)\right)\right) \\ &= m_m \left(G \frac{M_E}{R_E^2}\right) y \end{aligned}$$

All that is left is to realize that

$$\left(G \frac{M_E}{R_E^2}\right)$$

has units of acceleration. This is just g

$$g = \left(G \frac{M_E}{R_E^2}\right)$$

so we have

$$\Delta U = m_m g y$$

and there is no contradiction. But we should realize that this is an approximation. The more accurate version of our potential energy is

$$U_2 = -G \frac{M_E m_m}{R_2}$$

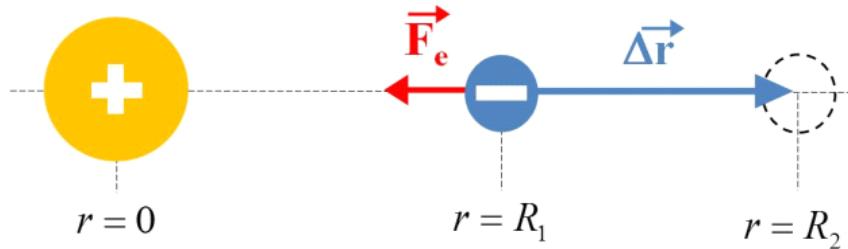
Likewise we should expect that for charges

$$\Delta U_C = -q_m E d$$

is an approximation that is only good when the field, E , can be approximated as a constant magnitude and direction and that the distribution of charge, q_m , is not spatially too big. With this understanding, we can understand electrical potential energy of point charges.

Point charges potential

Suppose we now take a positive charge and define it's position as $r = 0$ and place a negative mover charge near the positive charge.



The work it would take to move the charge a distance $\Delta r = R_2 - R_1$ would be

$$\Delta U_e = -W_e = - \int \vec{F}_e \cdot d\vec{r}$$

The magnitude of the electrical force is

$$F_e = \frac{1}{4\pi\epsilon_o} \frac{Q_E q_m}{r^2}$$

once again $\vec{F}_e \cdot d\vec{r} = -F_e dr$ and

$$\begin{aligned} \Delta U_e &= - \int_{R_1}^{R_2} \left(-\frac{1}{4\pi\epsilon_o} \frac{Q_E q_m}{r^2} \right) dr \\ &= \frac{Q_E q_m}{4\pi\epsilon_o} \int_{R_1}^{R_2} \frac{dr}{r^2} \end{aligned}$$

and we realize that this is exactly the same integral we faced in the gravitational case.

The answer must be

$$\Delta U_e = -\frac{1}{4\pi\epsilon_o} \frac{Q_E q_m}{R_2} + \frac{1}{4\pi\epsilon_o} \frac{Q_E q_m}{R_1}$$

The similarity is hardly a surprise since the force equation for the Coulomb force is really just like the force equation for gravity.

It makes sense to choose the zero point of the electric potential energy the same way we did for the gravitational potential energy since the equations is the same. We will

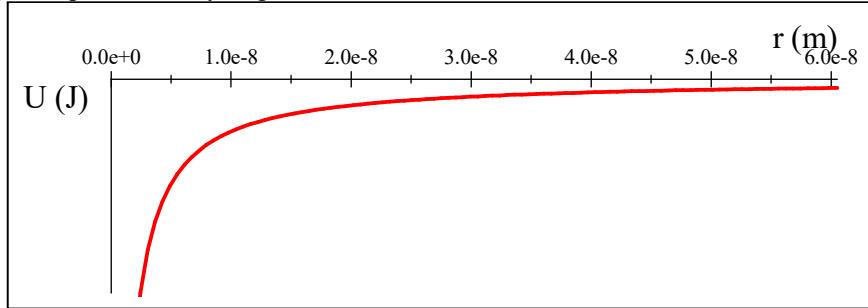
pick $U = 0$ at $r = \infty$. Then we expect that

$$U_e = -\frac{1}{4\pi\epsilon_0} \frac{Q_E q_m}{r}$$

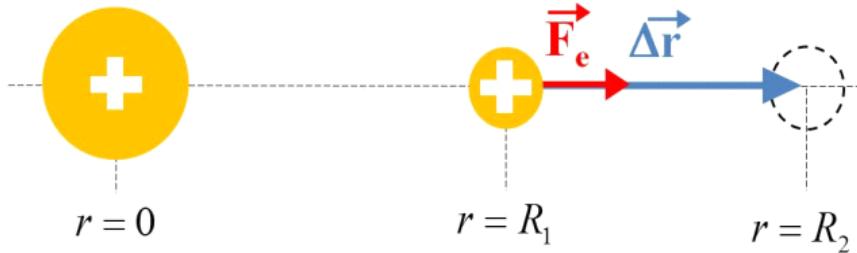
Question 223.29.4

is the electrical potential energy stored by having the charges in this configuration.

Again the negative sign shows that the two opposite charges will be bound together by the attractive force. Here is a graph of the electrical potential energy of an electron and a proton pair, like a Hydrogen atom.



Of course we remember that there is a large difference between electrical and gravitational forces. If the two charges are the same sign, then they will repel and the potential must be different for that situation. If we redraw our diagram for this case, we realize that the sign of the force must change.



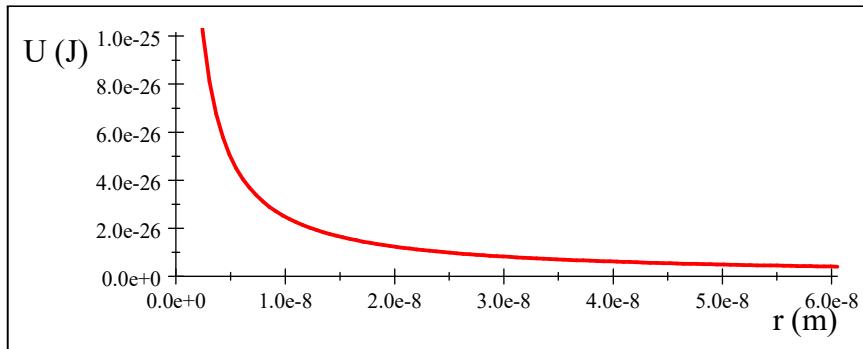
$$\Delta U_e = -W_e = - \int_{R_1}^{R_2} \left(+\frac{1}{4\pi\epsilon_0} \frac{Q_E q_m}{r^2} \right) dr$$

this will change all the signs in our solution

$$\Delta U_e = +\frac{1}{4\pi\epsilon_0} \frac{Q_E q_m}{R_2} - \frac{1}{4\pi\epsilon_0} \frac{Q_E q_m}{R_1}$$

then

$$U_e = +\frac{1}{4\pi\epsilon_0} \frac{Q_E q_o}{r}$$



Now we can see that the potential energy gets larger as the two like charges get nearer. It takes energy to make them get closer. This is clearly not a bound situation.

Three point charges.

Question 223.29.5

Suppose we have three like charges. What will the potential energy of the three-charge system be?

Let's consider the charges one at a time. If I move one charge, q_1 , from infinitely far away, there is no environmental electric field, so there is no force, since we need two charges for there to be a force. Then there is no potential energy. This is like a rock floating in deep space far away from anything else in the universe. It just sits there, there is no potential for movement, so no potential energy. But when we bring in another charge, q_2 , then q_1 is an environmental charge making a field and q_2 is our mover charge. Then q_2 will take an amount of work equal to

$$U_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}}$$

to move in the charge because the two charges repeal each other. There is a force, so now there is an amount of potential energy associated with the work done to move the charges together.

Suppose we had chosen to bring in the other charge, q_3 , instead. Charge q_1 forms an environmental field. It takes an amount of energy

$$U_{13} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}}$$

to bring in the third charge. But if the second charge were already there, the second charge also creates an environmental field, so it also creates a force on the third charge. So it will take more work to bring in the third charge.

$$U_3 = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}} + \frac{1}{4\pi\epsilon_0} \frac{q_2 q_3}{r_{23}}$$

So the total amount of work involved in bringing all three charges together

$$U = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} + \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}} + \frac{1}{4\pi\epsilon_0} \frac{q_2 q_3}{r_{23}}$$

then the potential energy difference would be

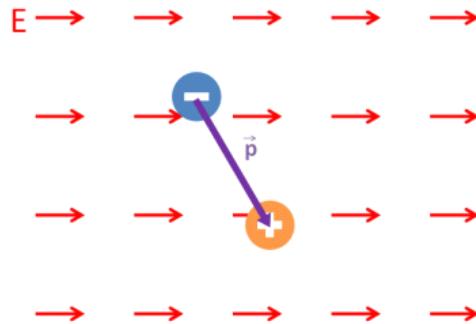
$$\begin{aligned}\Delta U &= U_f - U_i = -W \\ &= U_f - 0 \\ &= \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}} + \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{r_{13}} + \frac{1}{4\pi\epsilon_0} \frac{q_2 q_3}{r_{23}}\end{aligned}$$

which we can generalize as

$$U = \frac{1}{4\pi\epsilon_0} \sum_{i < j} \frac{q_i q_j}{r_{ij}}$$

for any number of charges. We simply add up all the potential energies. This is one reason to use electric potential energy in solving problems. The electric potential energies just add, and they are not vectors, so the addition is simple.

Dipole potential energy



Let's try out our new idea of potential energy for point charges on a dipole. We will try to keep this easy, so let's consider the dipole to be in a constant, uniform electric field. We know there will be no net force. The work done to move a charge we have stated to be

$$W = \int \vec{F}_e \cdot d\vec{r}$$

but in this case, we know the net force on the dipole is zero.

However, we can also do some work in rotating something

$$W_{rot} = \int \tau_e d\theta$$

we know from before that the magnitude of the torque is

$$\tau = pE \sin \theta$$

so

$$\begin{aligned} W_{rot} &= \int_{\theta_1}^{\theta_2} pE \sin \theta d\theta \\ &= pE (\cos \theta_1 - \cos \theta_2) \end{aligned}$$

this must give

$$\begin{aligned} \Delta U &= -W_{rot} = U_f - U_i \\ &= -pE (\cos \theta_2 - \cos \theta_1) \end{aligned}$$

then we can write as

$$U = -pE \cos \theta$$

This is the rotational potential energy for the dipole. We can write this as an inner product

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

What does this mean? It tells us that we have to do work to turn the dipole.

Let's go back to our example of a microwave oven. If the field is $E = 200 \text{ V/m}$, then how much work does it take to turn the water molecules?

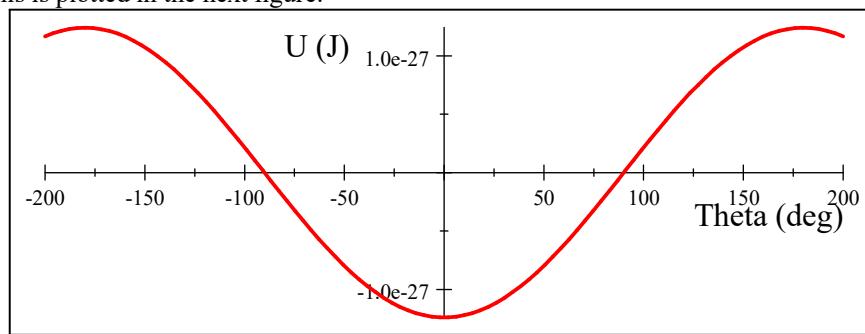
Remember that the dipole moment for a water molecule is something like

$$p_w = 6.2 \times 10^{-30} \text{ C m}$$

so we have

$$\begin{aligned} U &= -(6.2 \times 10^{-30} \text{ C m})(200 \text{ V/m}) \cos \theta \\ &= -1.24 \times 10^{-27} \text{ J} \cos \theta \end{aligned}$$

This is plotted in the next figure.



At zero degrees we can see that it takes energy (work) to make the dipole spin. It will try to stay at zero degrees and a small displacement from zero degrees will

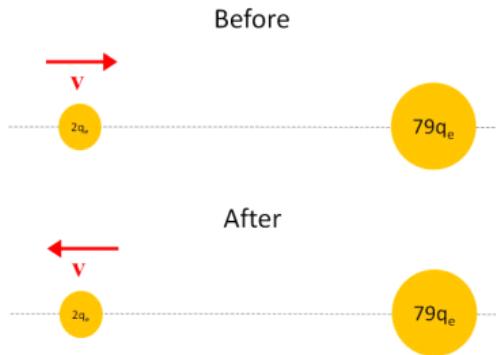
will cause the dipole to oscillate around $\theta = 0$ but it will return to $\theta = 0$ as the added energy is dissipated. Then $\theta = 0$ rad is a stable equilibrium. Conversely, at $\theta = \pi$ rad we are at a maximum potential energy. We get rotational kinetic energy if we cause any small displacement $\Delta\theta$. The dipole will angularly accelerate. $\theta = \pm\pi$ rad is an unstable equilibrium.

Shooting α -particles

Let's use electric potentials to think about a famous experiment. Ernest Rutherford shot α -particles, $q = +2q_e$ at gold nuclei, $q = +79q_e$. How close will the α -particles get if the collision is head-on and the initial speed of the α -particles is 3×10^6 m/s?

The easiest way to approach this is to use conservation of energy. The energies before and after must be the same because we have no frictional or dissipative forces.

The before and after pictures are as shown. The α -particle, of course, is our mover.



We can write

$$K_i + U_i = K_f + U_f$$

when the α -particles are at their closest distance to the gold nuclei, then $K_f = 0$. We can envision starting the α -particles from effectively an infinite distance away. Then $U_i \approx 0$, so

$$\frac{1}{2}m_\alpha v^2 = \frac{1}{4\pi\epsilon_0} \frac{Q_{Au}q_\alpha}{r}$$

Solving for r gives

$$\begin{aligned} r &= \frac{1}{4\pi\epsilon_0} \frac{\frac{1}{2}m_\alpha v^2}{Q_{Au}q_\alpha} \\ &= \frac{1}{2\pi\epsilon_0} \frac{(79q_e)(4q_e)}{m_\alpha v^2} \end{aligned}$$

then

$$\begin{aligned} r &= \frac{1}{2\pi \left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2} \right)} \frac{158 (1.602 \times 10^{-19} \text{ C})^2}{(6.6422 \times 10^{-27} \text{ kg}) (3 \times 10^6 \text{ m/s})^2} \\ &= 1.2198 \times 10^{-12} \text{ m} \end{aligned}$$

This is a very small number! and it sets a bound on how large the nucleus of the gold atom can be.

Next lecture, we will try to make our use of electrical potential energy more practical by defining the electrical potential energy per unit charge, and applying this to problems involving moving charges (like those in electric circuits).

Basic Equations

30 Electric Potentials

We defined electrical potential energy last time. We used an analogy with gravitational fields and gravitational potential energy. But there is a missing piece. The gravitational environment property

$$g = \left(G \frac{M_E}{R_E^2} \right)$$

(where here the subscript E is for the environmental object) showed up in our equation for the gravitational potential

$$U = - \left(G \frac{M_E}{R_E} \right) m_o$$

We found the same form for the electrical potential energy.

$$U_{12} = \frac{1}{4\pi\epsilon_o} \frac{q_1 q_2}{r_{12}}$$

or we could write this as

$$U_{12} = \left(\frac{1}{4\pi\epsilon_o} \frac{q_1}{r_{12}} \right) q_2$$

where charge q_2 would be our mover charge. By analogy, then

$$\frac{1}{4\pi\epsilon_o} \frac{q_1}{r_{12}}$$

must represent the environment set up by q_1 . And sure enough, it has a q_1 in it. But this does not have the units of electric field. So it must be a new quantity. We will need a name for this new representation of the environment created by q_1 .

Fundamental Concepts

- Electric potential is a representation of the electric field environment.
- Electric potential is defined as the potential energy per unit charge.
- Equipotential lines are drawn to show constant electric potential surfaces
- The volt as a measure of electric potential
- The electron-volt as a measure of energy (and speed).

Electric Potential Difference

Question 223.30.1

Let's give a symbol and a name to our new environment quantity.

$$V_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{12}}$$

where we understand that q_1 is making the environment and we are measuring that environment a distance r_{12} from q_1 . Thus q_1 is the environmental charge.

Then

$$\begin{aligned} U_{12} &= \left(\frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{12}} \right) q_2 \\ &= (V_{12}) q_2 \end{aligned}$$

It's traditional to drop the subscripts on the V

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{r}$$

where we understand that an environmental charge labeled just q is making the environment and q_2 is a distance r from q . In that case we can write

$$U_{12} = (V) q_2$$

or

$$V = \frac{U_{12}}{q_2}$$

This new environment representation appears to be an amount of potential energy per unit charge. In general any electrical potential energy (U) per unit charge (q) is called an *electric potential*.

$$V = \frac{U}{q}$$

This is a somewhat unfortunate name, because it sounds like electric potential energy. But it is not, it is a representation of the environment set up by the electric field. We don't get electric potential energy without multiplying by a charge. $U = Vq_o$.

We will give electric potential the symbol V but usually the important quantity is a change in potential energy, then

$$\Delta V = \frac{\Delta U}{q} \quad (30.1)$$

If I know ΔV for a configuration of charge (like our capacitor plates) then I can find the ΔU of different charges by multiplying by the amount of charge in each case

$$\Delta U_1 = q_1 \Delta V$$

$$\Delta U_2 = q_2 \Delta V$$

⋮

which is convenient if I am accelerating many different charges. We do this in linear accelerators or “atom smashers” so this is important to physicists! We can see that the

units of ΔV must be

$$\frac{\text{J}}{\text{C}} = \text{V} \quad (30.2)$$

which has been named the *Volt* and is given the symbol, V.

Now this may seem familiar. Can you think of anything that carries units of volts? Let's consider a battery. In our cell phones we have something like a 3.8 V lithium-ion battery. Inside the battery we would expect that a charge would experience a potential energy difference. We use the battery so we can convert that potential energy into some other form of energy (e.g. radio wave energy for our phone's wifi). The potential energy achieved depends on the charge carrier. We would have electrons in metals but we would have ions in a solution. This is so convenient to express the potential energy per unit charge, that it is the common form or expressing the energy given by most electrical sources.

Question 223.30.2

Question 223.30.3

Electric Potential

Let's write out the electric potential difference between points *A* and *B*. It is the change in potential energy per unit charge as the charge travels from point *A* to point *B*

$$\Delta V = V_B - V_A = \frac{\Delta U}{q} \quad (30.3)$$

This is clearly a measure of how the environment changes along our path from *A* to *B*.

Let's reconsider gravitational potential energy. We remember that if the field is uniform (that is, if we are near the Earth's surface so the field seems uniform) we can set the zero point of the potential energy anywhere we find convenient for our problem, with the provision that once it is set for the problem, we have to stick with our choice.

One logical choice for many electrical appliances is to set the Earth's potential equal to zero. Note! this is not true for point mass problems where we have already set the potential energy $U = 0$ at $r = \infty$.

In our gravitational analogy, this is a little bit like mean sea level. Think of river flow. The lowest point on the planet is not mean sea level. But any water above mean sea level will tend to flow downward to this point. Of course, if we have land below mean sea level, the water would tend to continue downward (like water flows to the Dead Sea). The direction of water flow is given by the potential energy difference, not the actual value of the potential energy. It is the same way with electric potential. If we have charge at a potential that is higher than the Earth's potential, then charge will flow toward the Earth.

Question 223.30.4

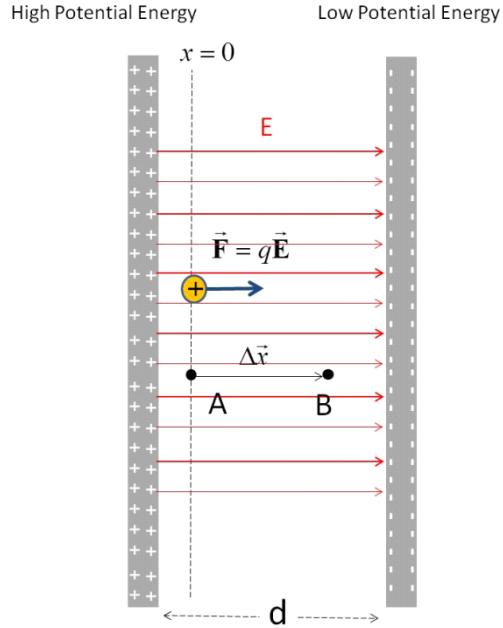
Consider a 9 V battery. If the negative terminal is connected to a grounding rod or metal water pipe, it will be at the electric potential of the Earth while its positive terminal will be at $\Delta V = 9$ V above the Earth's potential. Likewise, in your home, you probably have a 110 V outlet. One wire is likely set to the potential of the Earth by connecting it to a ground rod. The others are at $\Delta V = 110$ V above it²⁵.

In our phones, we don't have a ground wire, so we cannot guarantee that the negative terminal of the battery is at the same potential as the Earth. If our appliances in our house are not all grounded to the same potential, there is a danger that there will be a large enough difference in their potentials (think potential energy per unit charge) to cause the charges to accelerate from one appliance to another. It is the difference in potential that counts! This is a spark or shock that could hurt someone or damage equipment. That is why we now use grounded outlets. These outlets have a third wire that is tied to all the other outlet's third wire and also tied physically to the ground near your house or apartment. This way, all appliances are ensured to have the same low electric potential point.

Example, potential of a capacitor

Let's calculate the potential of our favorite device, the capacitor.

²⁵ House voltages are alternating voltages. We will deal with them later in this course.



The nice uniform field makes this a useful device for thinking about electric potentials.

We have found that field to be

$$E = \frac{\eta}{\epsilon_o}$$

with a direction from positive to negative. The work to push a mover charge from one side to the other is given by

$$W = \int F_e \cdot dx$$

The force is uniform since the field is uniform (near the middle at least)

$$F_e = q_o E$$

then our work becomes

$$W = \int q_o E \cdot dx$$

$$= q_o E \Delta x$$

and the amount of potential energy is

$$|\Delta U| = |-q_o E \Delta x|$$

We can set the zero potential energy point anywhere we want, but it is tradition to set $U = 0$ at the negative plate. If we do this we end up with the potential energy difference going from the negative plate to the positive plate being

$$\Delta U = q_o Ed$$

Then if we go from the negative plate to the positive plate we have a positive ΔU .

We have seen all this before when we compared the electric potential energy of a uniform gravitation field and a uniform electrical field. Now let's calculate the electric potential difference

$$\Delta V = \frac{\Delta U}{q_o} = \frac{q_o E d}{q_o} = Ed$$

Remember that the field is created by the charges on the capacitor plates, so it exists whether we put any q_o inside of the capacitor or not. Then the potential difference must exist whether or not there is a charge q_o inside the capacitor.

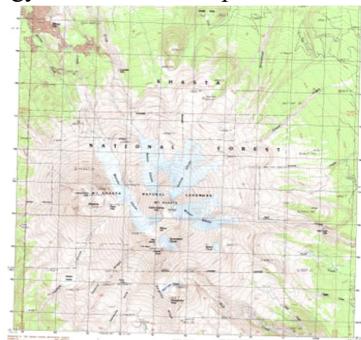
You probably already know that a voltmeter can measure the electric potential difference between two points, say, the plates of a capacitor. If we use such a meter we could find the field inside the capacitor (well, almost, remember our approximation is good for the center of the plates).

$$E = \frac{\Delta V}{d}$$

Equipotential Lines

Question 223.30.5

We need a way to envision this new environmental quantity that, like a field, has a value throughout all space. Our analogy with gravity gives us an idea. Suppose we envision the height potential energy as the top of a hill. Then the low potential energy would be the bottom of the hill. We know from our Young Men and Young Women's Camp experiences how to show a change in gravitational potential energy. We plot on a map lines of constant potential energy. We call it constant elevation, but since near the Earth's surface $U_g = mgh$ the potential energy is proportional to the height, so we can say these lines are lines of constant potential energy. Here is an example for Mt. Shasta.



Map courtesy USGS, Picture is in the Public Domain.

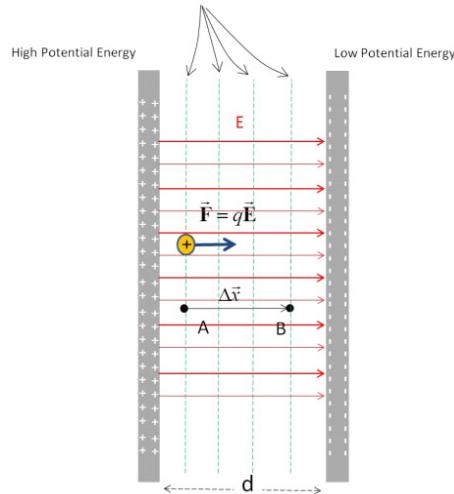
We can think of these lines of constant potential energy as paths over which the

gravitational field does no work. If we walked along one of these lines we would get neither higher nor lower and though we might do work to move us to overcome some friction, the gravitational field would do no work. And we would do no work in changing elevation.

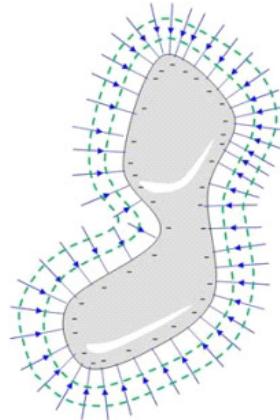
Question 223.30.6

Likewise we can draw lines of equal potential for our capacitor. When moving along these lines the electric field would do no work.

Equipotential Surfaces



Of course we could draw these lines for a crazier device. Say, for our charged conductor



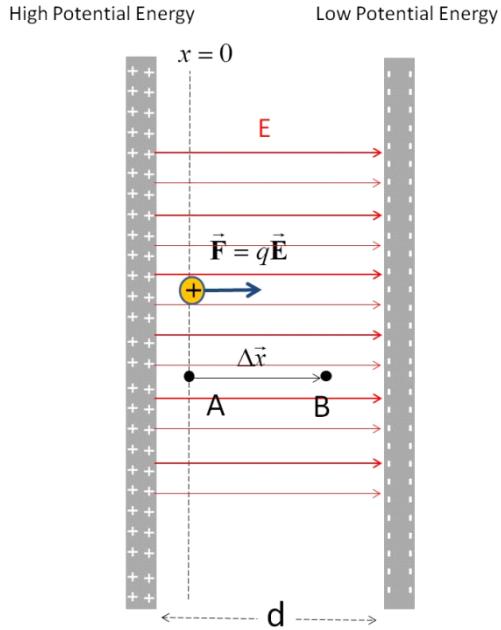
Notice that our equal potential lines are always perpendicular to the field. From

$$W = \int q_o \vec{E} \cdot d\vec{x}$$

we can see that if the path we travel is perpendicular to the field, no work is done. This is like us marching along around the mountain neither going up nor down.

Electron Volt

Suppose I set up our uniform electric field device again



We are not including any gravitational field, so the directions involved are all relative to the placement of the capacitor plate orientation.

This time, suppose I make the potential difference $\Delta V = 1$ V. I release a proton near the high potential side. What is the kinetic energy of the proton as it hits the low potential side? From the work energy theorem

$$W_{nc} = \Delta K + \Delta U$$

and if we do this in a vacuum so there is no non-conservative work,

$$\Delta K = -\Delta U$$

$$K_f - K_i = -\Delta U$$

$$K_f = -\Delta U$$

We can find the potential energy loss from what we just studied

$$\Delta V = \frac{\Delta U}{q}$$

so we can find the potential energy as

$$\Delta U = q\Delta V$$

but remember we are going from a high to a low potential

$$\Delta V = V_f - V_i$$

this will be negative, so the potential energy change will be negative too.

$$\begin{aligned} K_f &= -\Delta U \\ &= -q\Delta V \end{aligned}$$

which will be a positive value (which is good, because I don't know what negative kinetic energy would mean).

$$K_f = -q\Delta V$$

We can find the amount of energy in Jules

$$\begin{aligned} K_f &= -(1.6 \times 10^{-19} \text{ C})(-1 \text{ V}) \\ &= 1.6 \times 10^{-19} \text{ J} \end{aligned}$$

since we defined a volt as $V = \frac{J}{C}$.

You might think this is not very useful, but remember that $K = \frac{1}{2}mv^2$. The kinetic energy is related to how fast the proton is going. In a way, the kinetic energy tells us how fast the particle is going (we know its mass). If you read about the Large Hadron Collider at CERN, in Switzerland the "speeds" of the particles will be given in energy units that are multiples of 1.6×10^{-19} J. We call this unit an electron-volt (eV).



Beam magnet and Section of the Beam Pipe of the LHC. This section is actually no longer used and is in a service area 100 m above the operating LHC. The people you see are part of a BYU-I Physics Department Tour of the facility.

We can finish this problem by finding the speed of the particle

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

so

$$\frac{2K}{m} = v^2$$

or

$$\begin{aligned} v &= \sqrt{\frac{2K}{m}} \\ &= \sqrt{\frac{2(1.6 \times 10^{-19} \text{ J})}{1.00728 \text{ u} \frac{1.6605 \times 10^{-27} \text{ kg}}{1 \text{ u}}}} \\ &= 13832 \frac{\text{m}}{\text{s}} \end{aligned}$$

Which is pretty fast, but the Large Hadron Collider at CERN can provide energies up to 7×10^{14} eV which would give our proton a speed of 99.9999991% of the speed of light.



CERN CMS detector during a maintenance event. The bright metal pipe seen in the middle of the detector is the beam pipe through which the accelerated protons travel.

Note the workers near the scaffolding for scale.

Note that this energy would seem to provide a faster speed—faster than light! But with energies this high we have to use Einstein's theory of Special Relativity to calculate the particle speed. And, sadly, that is not part of this class. If you are planning to work on the GPS system, or future space craft, you might need to take yet another physics class so you can do this sort of calculation.

You might guess that we will want to know the electric potential of more complex configurations of charge. We will take on this job in the next lecture.

Basic Equations

The electric potential is the electrical potential per unit charge

$$\Delta V = V_B - V_A = \frac{\Delta U}{q}$$

For the special case of a constant electric field in a capacitor the electrical potential is just

$$\Delta V = E\Delta s$$

where Δs is the distance traveled from one side of the capacitor to the other.

The unit

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

31 Electric potential of charges and groups of charges

Now that we have a new representation of the environment created by environmental charges, we will need to be able to calculate values for that representation for different configurations of charge like we did for electrical fields. But there is a huge benefit in using the electric potential representation, electric potentials are not vectors! So we don't have to deal with the vector nature of the field environment. The vector nature is still there, but we will ignore it. This means we will give up being able to give up vector directions for movement of our mover charges in many cases. But we can know much about the movement and the equations will be much simpler. We will take on the usual cases of environments from a point charge, a collection of point charges, and a continuous distribution of charges.

Fundamental Concepts

- Finding the electric potential of a point charge
- Finding the electric potential of two point charges
- Finding the electric potential of many point charges
- Finding the electric potential of continuous distributions of point charges.

Point charge potential

The capacitor was an easy electric potential to describe. Let's go back to a slightly harder one, the potential due to just one point charge. The potential energy depends on two charges

$$U_e = -\frac{1}{4\pi\epsilon_0} \frac{Qq}{r}$$

but the potential just depends on one.

$$V = \frac{U}{q}$$

where U is a function of q , so a charge will cancel. But which charge do we divide by?

We need two charges to make a force,

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}$$

but when we defined the electric field we said the field from charge 1 would be there whether or not charge 2 was present. The situation is the same for electric potential.

We say we have an electric potential due to the first charge even if the second charge is not there. This is like saying there is a potential energy per unit rock, even if there is no rock to fall down the hill. The hill is there whether or not we are throwing rocks down it.

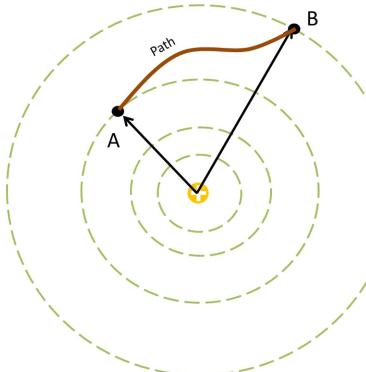
For electric potential, the potential is due to the field, and the field is there whether another charge is there or not.

Let's find this potential due to just one charge, but let's find it in a way that demonstrates how to find potentials in any situation. After all, from what we know about point charges, we can predict that

$$V = \frac{U}{q} = \frac{\frac{1}{4\pi\epsilon_0} \frac{Qq}{r}}{q} = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$

But not all situations come so easily. We only know forms for U for capacitors and point charges so far. So let's see how to do this in general, and compare our answer for the point charge with what we have guessed from knowing U .

Symmetry tells us the field will be radial, so the equipotential surfaces must be concentric spheres. Here is our situation:



We wish to follow the marked path from A to B finding the potential difference $\Delta V = V_B - V_A$.

Remember that the field due to a charge q is radially outward from the charge. To find

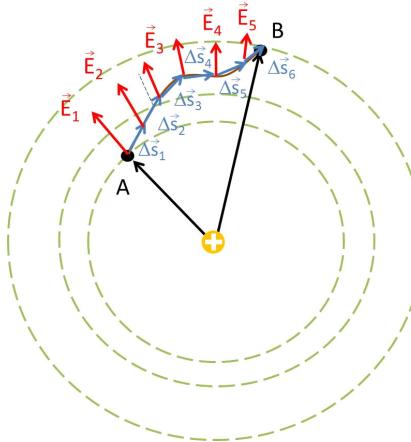
the potential we start with what we found last lecture, for a constant field

$$\Delta V = \frac{\Delta U}{q_o} = \frac{q_o E \Delta s}{q_o} = E \Delta s$$

where s is the path length along our chosen path from A to B . For our capacitor, this was just the distance from one side to the other, but here we need to be more general. We should really write this as

$$\Delta V = \vec{E} \cdot \vec{\Delta s}$$

Further, our field, E , changes, so technically this value for ΔV is not correct. But if we take very small paths, $\Delta \vec{s}$, then the field will be nearly constant over the small distances. Then we can add up the contribution of each small distance, $\Delta \vec{s}_i$ to deal with the entire path from A to B for our point charge geometry.



That is, we take a small amount of path difference $\Delta \vec{s}_i$ and add up the contribution, $\vec{E} \cdot \vec{\Delta s}_i$ from this small path. Then we can repeat this for the next $\Delta \vec{s}_{i+1}$ and the next, until we have the contribution of each piece of the path. We can call the contribution from one piece.

$$\Delta V_i = \vec{E} \cdot \vec{\Delta s}_i$$

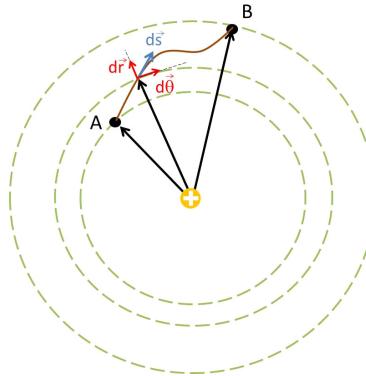
The total potential difference would be

$$\Delta V = \sum_i \vec{E} \cdot \vec{\Delta s}_i$$

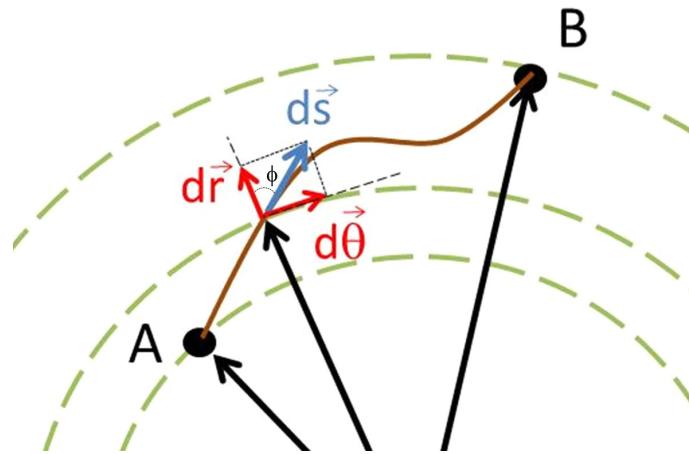
In the limit that the Δs_i become very small this becomes an integral

$$\Delta V = - \int_A^B \vec{E} \cdot d\vec{s} \quad (31.1)$$

where A and B are any two points.



Here is an expansion of the region about A and B .



Let's divide up our $d\vec{s}$ into components in the radial and azimuthal directions

$$d\vec{s} = (dr\hat{r} + rd\theta\hat{\theta})$$

from trigonometry we can see that

$$\cos \phi = \frac{dr}{ds}$$

and

$$\sin \phi = \frac{d\theta}{ds}$$

so

$$dr = ds \cos \phi$$

$$d\theta = ds \sin \phi$$

and we can write

$$d\vec{s} = (ds \cos \phi \hat{r} + rds \sin \phi \hat{\theta})$$

The field due to the point charge is

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \quad (31.2)$$

if we take

$$\begin{aligned}\vec{E} \cdot d\vec{s} &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \cdot d\vec{s} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \cdot (ds \cos \phi \hat{r} + ds \sin \phi \hat{\theta})\end{aligned}$$

we get only a radial contribution since $\hat{r} \cdot \hat{\theta} = 0$. Then

$$\begin{aligned}\vec{E} \cdot d\vec{s} &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \cdot ds \cos \phi \hat{r} + 0 \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} ds \cos \phi\end{aligned}$$

where ϕ is the angle between $d\vec{s}$ and \hat{r} and where we recall that $\hat{r} \cdot \hat{r} = 1$. Recalling that

$$dr = ds \cos \phi$$

we can eliminate ϕ from our equation

$$\vec{E} \cdot d\vec{s} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} dr$$

and we can integrate this!

$$\begin{aligned}\Delta V &= - \int_{r_A}^{r_B} \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} dr \\ &= - \frac{q}{4\pi\epsilon_0} \int_{r_A}^{r_B} \frac{1}{r^2} dr \\ &= \frac{q}{4\pi\epsilon_0} \frac{1}{r} \Big|_{r_A}^{r_B}\end{aligned}$$

so

$$\begin{aligned}\Delta V &= \frac{q}{4\pi\epsilon_0} \left(\frac{1}{r_B} - \frac{1}{r_A} \right) \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r_B} - \frac{1}{4\pi\epsilon_0} \frac{q}{r_A} \\ &= V_B - V_A\end{aligned}$$

Question 223.31.1

Note that the potential depends only on the radial distances from the point charge—not the path. We would expect this for conservative fields (where energy is conserved).

We know that, like potential energy, we may choose our zero point for the electric potential. For a point charge, we said we would take the $r_A = \infty$ point as $V = 0$.²⁶ So you will often see the potential for the point charge written as just

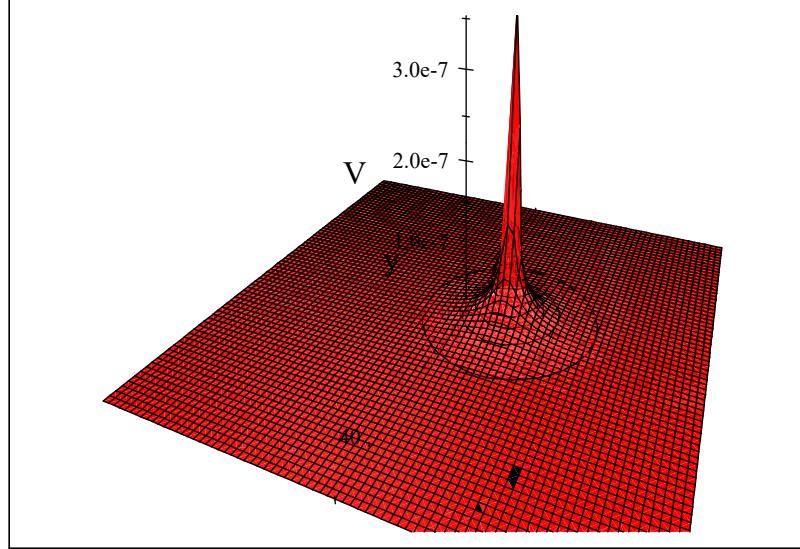
$$\Delta V = \frac{1}{4\pi\epsilon_0} \frac{q}{r_B}$$

²⁶ Remember this is because $U \rightarrow 0$ when $r \rightarrow \infty$.

or simply as

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{r} \quad (31.3)$$

Here is a plot of this with $q = 2 \times 10^{-9}$ C and the charge placed right at $x = 10$ m.



It is probably a good idea to state that in common engineering practice we kind of do all this backwards. We usually say we will charge up something until it has a particular voltage. This is because we have batteries or power supplies that are charge delivery services. They can provide enough charge to make some object have the desired voltage. By “desired voltage” we always mean the voltage at a conductor surface in our apparatus.

Early *electrodes* were spherical, so let’s consider making a spherical conductor have a particular potential at its surface. A sphere of charge with radius R would have

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{R}$$

at its surface. We can guess this because Gauss’ law tells us that the field of a charged sphere is the same as that of a point charge with the same Q . Then it takes

$$Q = 4\pi\epsilon_0 RV$$

to get the voltage we want. The battery or power supply must provide this. If the power supply or battery has a large amperage (ability to supply charge) this happens quickly. But away from the electrode the potential falls off. We can find how it falls off by again

using

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$

but with charge

$$Q = 4\pi\epsilon_0 RV_o$$

so that

$$V = \frac{1}{4\pi\epsilon_0} \frac{4\pi\epsilon_0 RV_o}{r}$$

or

$$V = \frac{R}{r} V_o$$

where V_o is the voltage at the surface. We can see that as r increases, V decreases.

Two point charges

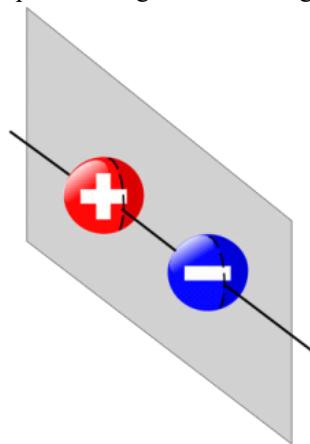
Question 223.31.2

Question 223.31.3

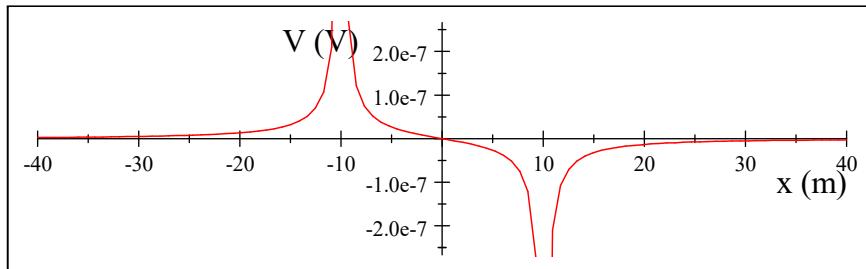
We can guess from our treatment of the potential energy of two point charges that the electric potential of two point charges is just the sum of the individual point charge potentials.

$$\begin{aligned} V &= V_1 + V_2 \\ &= \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_1} + \frac{1}{4\pi\epsilon_0} \frac{q_2}{r_2} \\ &= \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_1} + \frac{q_2}{r_2} \right) \end{aligned}$$

It is instructive to look at the special case of two opposite charges (our dipole). We can plot the electric potential in a plane through the two charges.



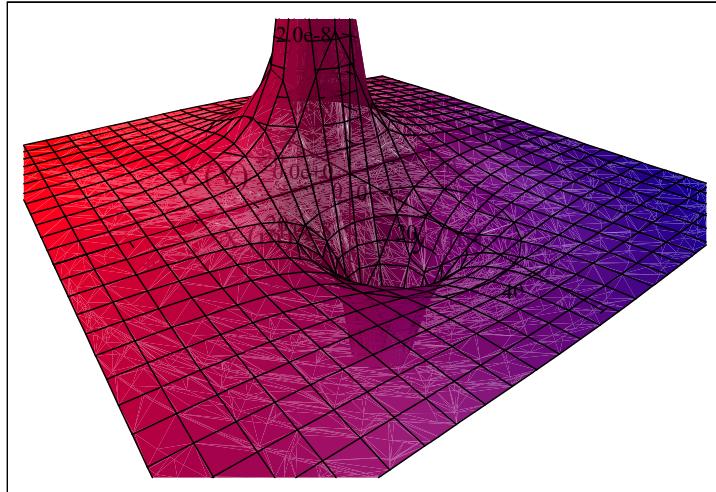
It would look like this



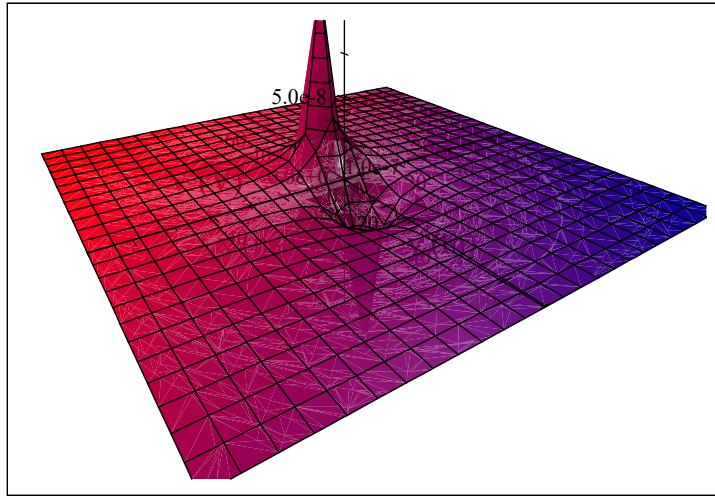
The charges ($q = 2 \times 10^{-9}$ C) were placed right at $x = \pm 10$ m. The potential

$$V = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_1} + \frac{q_2}{r_2} \right)$$

becomes large near $r_1 = R_o$ or $r_2 = R_o$ where R_o is the charge radius (which is very small, since these are point charges). Plotting the potential in two dimensions is also interesting. We see that near the positive charge we have a tall mountain-like potential and near the negative charge we have a deep well-like potential.



Notice the equipotential lines. The more red peak is the positive charge (hill), the more blue the negative charge (valley). A view from farther away looks like this



Of course the hill and the valley both approach an infinity at the point charge because of the $1/r$ dependence.

Lots of point charges

Question 223.31.4

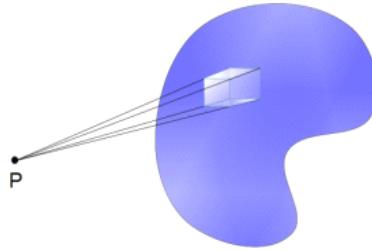
Suppose we have many point charges. What is the potential of the group? We just use superposition and add up the contribution of each point charge

$$V = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{r_i} \quad (31.4)$$

where r_i is the distance from the point charge q_i to the point of interest (where we wish to know the potential). Note that this is easier than adding up the electric field contributions. Electric potentials are not vectors! They just add as scalars.

Potential of groups of charges

Suppose we have a continuous distribution of charge. Of course, this would be made of many, many point charges, but if we have so many point charges that the distance between the individual charges is negligible, we can treat them as one continuous thing. If we know the charge distribution we can just interpret the distribution as a set of small amounts of charge dq acting like point charges all arranged into some shape.



Then for each charge dq we will have a small amount of potential

$$dV = \frac{1}{4\pi\epsilon_0} \frac{dq}{r} \quad (31.5)$$

and the total potential at some point will be the summation of all these small amounts of charge

$$V = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r} \quad (31.6)$$

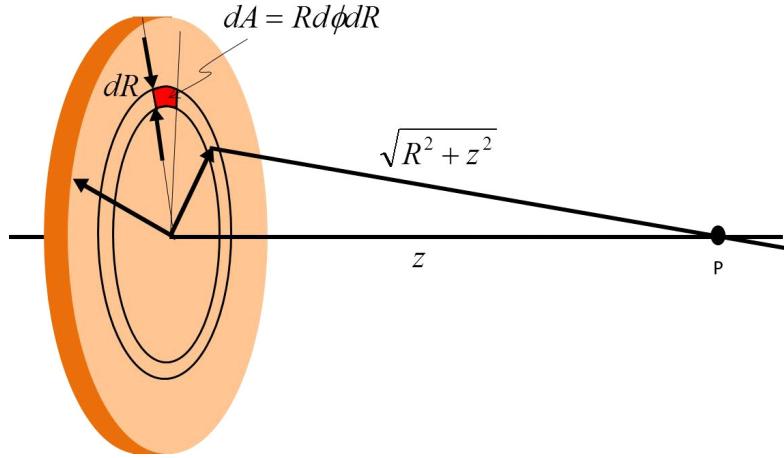
This looks a little like our integral for finding the electric field from a configuration of charge, but there is one large difference. There is no vector nature to this integral. So our procedure will have one less step

- Start with $V = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r}$
- find an expression for dq
- Use geometry to find an expression for r , the distance from the group of charges, dq , and the point P
- Solve the integral

Let's try one together

Electric potential due to a uniformly charged disk

We have found the field due to a charged disk. We can use our summation of the potential due to small packets of charge to find the electric potential of an entire charged disk.



Suppose we have a uniform charge density η on the disk, and a total charge Q , with a disk radius a . We wish to find the potential at some point P along the central axis.

To do this problem let's divide up the disk into small areas, dA each with a small amount of charge, dq . The area element is

$$dA = Rd\phi dR$$

so the charge element, dq , is

$$dq = \eta R d\phi dR$$

For each dq we have a small part of the total potential. The variable r is the distance from our small group of charges that we called dq to the point P . Then $r = \sqrt{R^2 + z^2}$ and our integral becomes

$$\begin{aligned} V &= \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r} \\ &= \frac{1}{4\pi\epsilon_0} \int \int \frac{\eta R d\phi dR}{\sqrt{R^2 + z^2}} \end{aligned}$$

We will integrate this. We will integrate over r from 0 to a and ϕ from 0 to 2π which will account for all the charge on the disk, and therefore all the potential.

Question 223.31.5

$$\begin{aligned}
 V &= \frac{1}{4\pi\epsilon_0} \int_0^{2\pi} \int_0^a \frac{\eta R d\phi dR}{\sqrt{R^2 + z^2}} \\
 &= \frac{\eta 2\pi}{4\pi\epsilon_0} \int_0^a \frac{R dR}{\sqrt{R^2 + z^2}} \\
 &= \frac{\eta 2\pi}{4\pi\epsilon_0} \sqrt{R^2 + z^2} \Big|_0^a \\
 &= \frac{\eta 2\pi}{4\pi\epsilon_0} \sqrt{a^2 + z^2} - \frac{\eta 2\pi}{4\pi\epsilon_0} z
 \end{aligned} \tag{31.7}$$

so

$$V = \frac{\eta}{2\epsilon_0} \left(\sqrt{a^2 + z^2} - z \right) \tag{31.8}$$

This is the potential at point P .

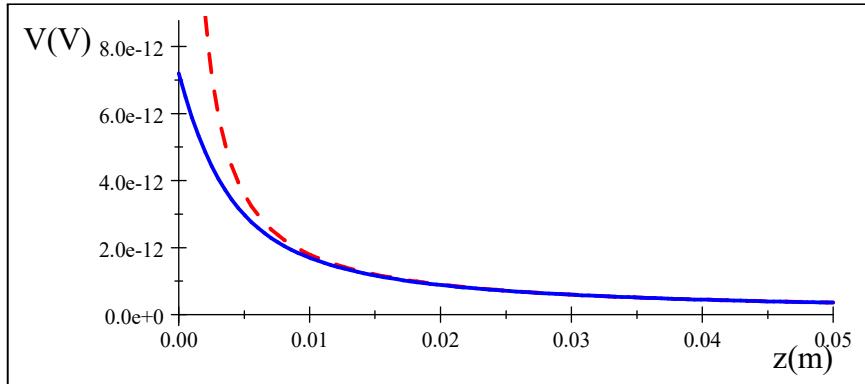
We compared our electric field solutions with the solution for a point charge. We can do the same for electric potentials. We can compare our solution to a point charge potential for an equal amount of charge. Far away from the disk, we expect the two potentials to look the same. The point charge equation is

$$V = \frac{Q}{4\pi\epsilon_0} \frac{1}{z}$$

Our disk gives

$$V = \frac{Q}{4\epsilon_0\pi} \frac{2}{a^2} \left(\sqrt{a^2 + z^2} - z \right) \tag{31.9}$$

They don't look much alike! But plotting both yields



The dashed line is the point charge, the solid line is our disk with a radius of 0.05 m and a total charge of 2 C. This shows that far from the disk the potential is like a point charge, but close the two are quite different as we would expect. This is a reasonable result.

We will calculate the potential due to several continuous charge configurations.

But, you may ask, since we knew the field for the disk of charge, couldn't we have found the electric potential from our equation of the field? We will take up this question in the next two lectures.

Basic Equations

The electric potential of a point charge is given by

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$

where the zero potential point is set at $r = \infty$.

Electric potentials simply add, so the potential for a collection of point charges is just

$$V = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{r_i}$$

To find the potential due to a continuous distribution of charge we use the following procedure:

- Start with $V = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r}$
- find an expression for dq
- Use geometry to find an expression for r
- Solve the integral

Since electric fields and electric potentials are both representations of the environment created by the environmental charge, there must be a way to calculate the potential from the field and *vice versa*. It will take us two lectures to do both.

32 Connecting potential and field

Fundamental Concepts

- The potential and the field are manifestations of the same physical thing
- We find the potential from the field using $\Delta V = - \int \vec{E} \cdot d\vec{s}$
- Fields and potentials come from separated charge

Finding the potential knowing the field

It is time to pause and think about the meaning of this electric potential. Let's trace our steps backwards. We defined the electric potential as the potential energy per unit charge:

$$\Delta V = \frac{\Delta U}{q}$$

where q is our mover and ΔV is a measure of the change in the environment between two points r_1 and r_2 measured from the environmental charge. ΔU is the change in potential energy as q moves. But the potential energy change is equal to the negative of the amount of work we have done in moving q

$$\Delta V = \frac{-W}{q}$$

which is equal to

$$\Delta V = \frac{-1}{q} \int \vec{F} \cdot d\vec{s}$$

where again $d\vec{s}$ is a general path length. But this force was a Coulomb force, which we know is related to the electric field

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

so we may rewrite the potential as

$$\begin{aligned}\Delta V &= - \int \frac{\vec{F}}{q} \cdot d\vec{s} \\ &= - \int \vec{E} \cdot d\vec{s}\end{aligned}$$

Question 223.32.3

which we found last lecture by analogy with our capacitor potential. Our line of reasoning in this lecture has been more formal, but we arrive at the same conclusion—**and it is an important one!** If we add up the component of field magnitude times the displacement along the path take from r_1 to r_2 we get the electric potential (well, minus the electric potential).

The electric field and the electric potential are not two distinct things. They are really different ways to look at the same thing—and that thing is the environment set up by the environmental charge. It is tradition to say the electric field is the principal quantity. This is because we have good evidence that the electric field *is* something. That evidence we will study at the end of these lectures, but in a nutshell it is that we can make waves in the electric field. If we can make waves in it, it must be something!²⁷

in our gravitational analogy, the gravitational field is the real thing. Gravitational potential energy is a result of the gravitational field being there. The change in potential energy is an amount of work, and the gravitational force is what does the work. No force, no potential energy. The gravitational field makes that force happen.

It is the same for our electrical force. The electrical potential is due to the Coulomb force, and the Coulomb force exists because the electric field is there.

If the field and the potential are really different manifestations of the same thing, we should be able to find one from the other. We have one way to do this. We can find the potential from the field, but we should be able to find the field from the potential. We will practice the first

$$\Delta V = - \int \vec{E} \cdot d\vec{s}$$

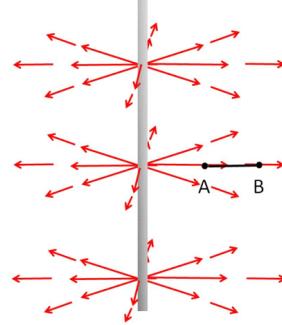
today, and then introduce how to find the field from the potential next lecture.

Finding the potential from the field.

Actually we did an example last lecture. We found the field of a point charge. But let's

²⁷ By the end of these lectures, we will try to make this a more convincing (and more mathematical) statement!

take on some harder examples in this lecture.



Let's calculate the electric potential due to an infinite line of charge. This is like the potential due to a charged wire. We already found the field due to an infinite line of charge

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{r} \hat{\mathbf{r}}$$

so we can use this to find the potential difference.

$$\Delta V = - \int_A^B \overrightarrow{\mathbf{E}} \cdot d\overrightarrow{s}$$

We need $d\overrightarrow{s}$. Of course $d\overrightarrow{s}$ could be in any direction. We can take components in cylindrical coordinates

$$d\overrightarrow{s} = dr\hat{\mathbf{r}} + rd\theta\hat{\theta} + dz\hat{\mathbf{z}}$$

Putting in our field gives

$$\begin{aligned} \Delta V &= - \int_A^B \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{r} \hat{\mathbf{r}} \cdot (dr\hat{\mathbf{r}} + rd\theta\hat{\theta} + dz\hat{\mathbf{z}}) \\ &= - \frac{2\lambda}{4\pi\epsilon_0} \int_A^B \frac{dr}{r} \end{aligned}$$

which we can integrate

$$\begin{aligned} \Delta V &= \left(-\frac{1}{2\pi\epsilon_0} \frac{\lambda}{r} \ln r_B - \left(-\frac{1}{2\pi\epsilon_0} \frac{\lambda}{r} \ln r_A \right) \right) \\ &= - \frac{1}{2\pi\epsilon_0} \frac{\lambda}{r} (\ln r_B - \ln r_A) \end{aligned}$$

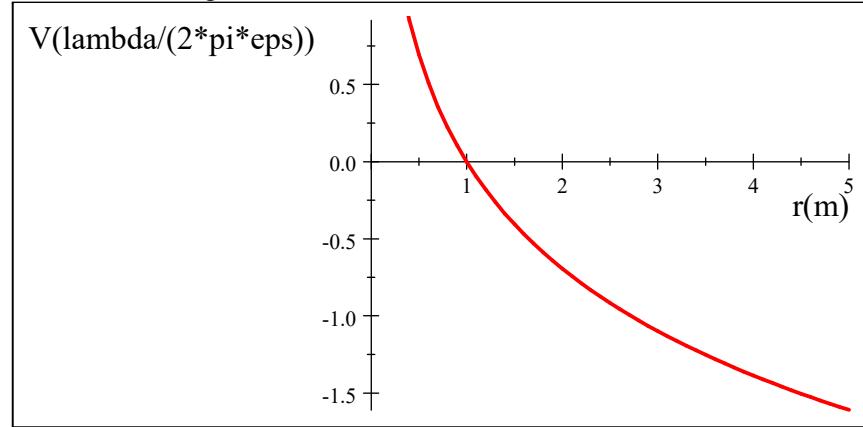
This example gives us a chance to think about our simple geometries and to consider when they are reasonable approximations to real charged objects. So long as neither r_A nor r_B are infinite, this result is reasonable. But remember what it looks like to move away from an infinite line of charge. No matter how far away we go, the line is still infinite. So we never get very far away. The terms

$$V_A = \frac{1}{2\pi\epsilon_0} \frac{\lambda}{r} (\ln r_A)$$

or

$$V_B = \frac{1}{2\pi} \frac{\lambda}{\epsilon_0} (\ln r_B)$$

would look something like this



The curve is definitely not approaching zero as r gets large. No matter how far we get from an infinite line of charge, we really never get very far compared with its infinite length. So the potential is not going to zero!

Our solution is good only when r_A and r_B are much smaller than the length of the line, that is, when our simple geometry is a good representation for something that is real, in this case, a finite length wire. But for $r_A, r_B \ll L$ this works.

We should also pause to think of the implications of this result for electronic equipment design. Our result means that adjacent wires in a cable or on a circuit board will feel a potential due to their neighbors—something we have to take into consideration in the design to ensure your equipment will work! This is one reason why we use shielded cables for delicate instruments, and for data lines, etc.

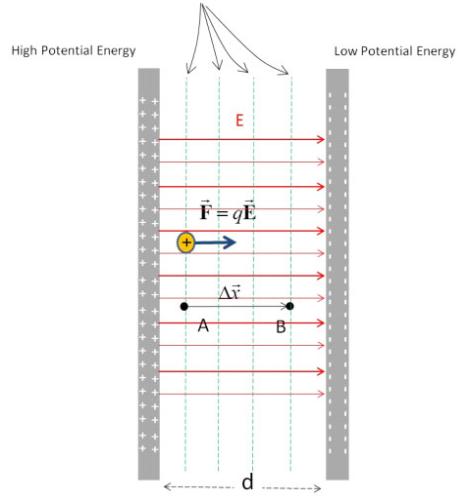
As a second example, let's tackle our friendly capacitor problem again. What is the potential difference as we cross the capacitor from point A to point B ? We already know the answer

$$\Delta V = Ed$$

But when we found this before, we assumed we knew the potential energy. This time let's practice using

$$\Delta V = - \int_A^B \vec{E} \cdot d\vec{s}$$

Equipotential Surfaces



We know the field is

$$E = \frac{\eta}{\epsilon_0}$$

so

$$\begin{aligned} \Delta V &= - \int_A^B \vec{E} \cdot d\vec{s} \\ &= - \int_A^B \frac{\eta}{\epsilon_0} ds \cos \theta \end{aligned}$$

where θ is the angle between the field direction and our $d\vec{s}$ direction. We could write

$$dx = ds \cos \theta$$

Then

$$\begin{aligned} \Delta V &= - \frac{\eta}{\epsilon_0} \int_A^B dx \\ &= - \frac{\eta}{\epsilon_0} (x_B - x_A) \\ &= - \frac{\eta}{\epsilon_0} \Delta x \end{aligned}$$

This is just

$$\Delta V = -E \Delta x$$

if we consider the negative side to be the zero potential, and we cross the entire capacitor, then

$$\begin{aligned} \Delta V &= -E (x_B - x_A) \\ &= -E (0 - d) \\ &= Ed \end{aligned}$$

as we expect. Note that we can now see how the positive result comes from our choice

of the zero voltage point.

Sources of electric potential

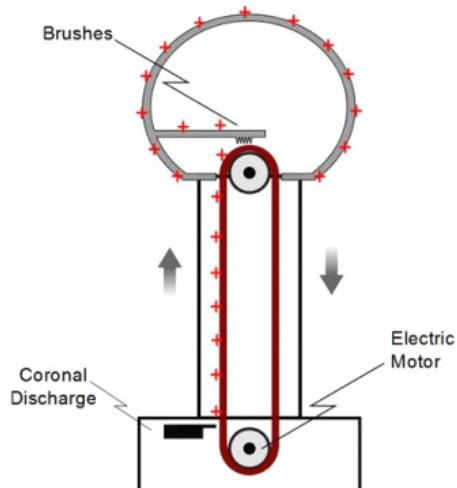
Question 223.32.4

We know that the electric potential comes from the electric field. And if we think about it, we know where the electric field comes from, charge. But we have found that equal amounts of positive and negative charge produce no net field. So normal matter does not seem to have any net electric field because the protons and electrons create oppositely directed fields, with no net result.

But if we separate the positive and negative charges, we do get a field. This is the source of all electric fields that we see, and therefore all electric potentials are due to separated charge.

We have used charge separation devices already in our lectures. Rubbing a rubber rod with rabbit fur transfers the electrons from the fur to the rod. Some of the charges that were balanced in the fur are now separated. So there is an electric field that creates an electric force. Then there must be an electric potential, since the potential is just a manifestation of the field.

We have also used a van de Graaff generator. It is time to see how this works.



In the base of the van de Graaff, there is a small electrode. It is charged to a large voltage, and charge leaks off through the air to a rubber belt that is very close. The rubber belt is connected to a motor. The motor turns the belt. The extra charge is stuck

on the belt, since the belt is not a conductor. The charge is carried up to the top where there is a large round electrode. A conducting brush touches the rubber belt, and the charge is able to escape the belt through the conductor. The charge spreads over the whole spherical electrode surface.

The belt keeps providing charge. Of course the new charge is repelled by the charge all ready accumulated on the spherical electrode, so we must do work to keep the belt turning and the charge ascending to the ball at the top. This is a mechanical charge separation device. It can easily build potential differences between the spherical top and the surrounding environment (including you) of 30000 V.

Much larger versions of this device are used to accelerate sub atomic particles to very high speeds.

Electrochemical separation of charge

Question 223.32.5

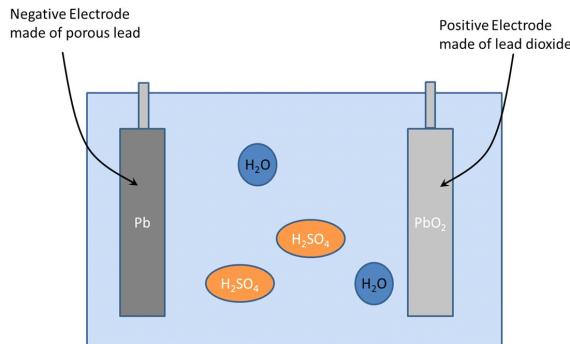
When you eat table salt, the NaCl ionic bond splits when exposed to polar water molecules, leaving a positively charged Na ion and a negatively charged Cl ion. This is very like the “bleeding” of charge from our charged balloons that we talked about earlier. We already know that the water molecules are polar, and the mostly positive hydrogens are attracted to the negatively charged Cl ions. This causes a sort of tug-o-war for the Cl ions. The positively charged Na ions pull with their coulomb force, and so do the positively charged hydrogens of the water molecules. If we have lots of water molecules, they win and the NaCl is broken apart. Water molecules are polar, but overall neutral. But now, with the Na and Cl ions, we have separated charge. We can make this charge flow, so we can get electric currents in our bodies. Our nervous system uses the positively charged Na ions to form tiny currents into and out of neurons as part of how nerve signaling works. Of course, NaCl is a pretty simple molecule. We could use more complex chemical reactions to separate charge.

batteries and emf

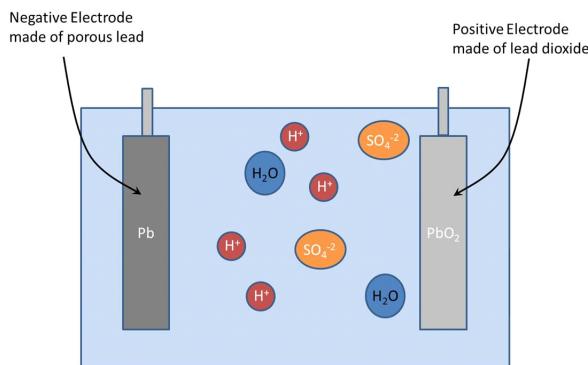
Most of us don't have a van de Graaff generator in our pockets. But most of us do have a charge separation device that we carry around with us. We call it a battery. But what does this battery do?

Somehow the battery supplies positive charge on one side and negative charge on the

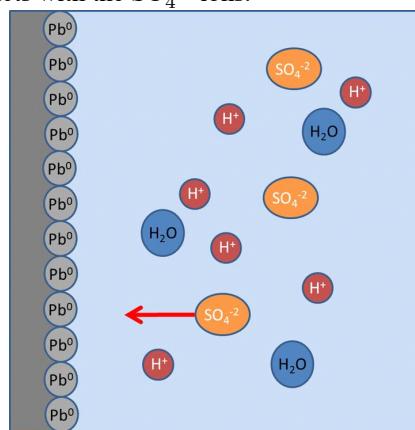
other side. This is accomplished by doing work on the charges. A lead acid battery is often used in automobiles. The battery is made by suspending two lead plates in a solution of sulfuric acid and water.



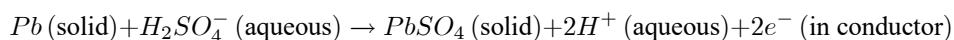
One plate is coated with lead dioxide. There is a chemical reaction at each plate. The sulfuric acid (H_2SO_4) splits into two H^+ ions and an SO_4^{2-} ion.



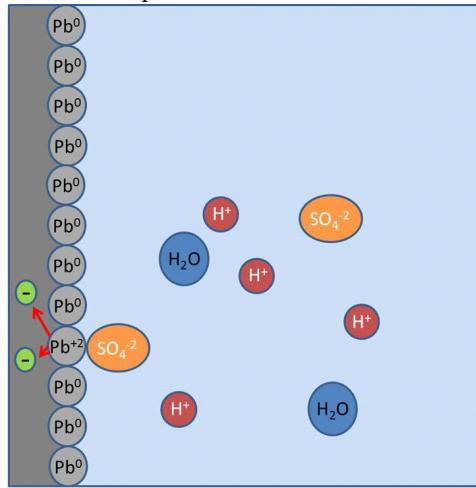
The plain lead plate reacts with the SO_4^{2-} ions.



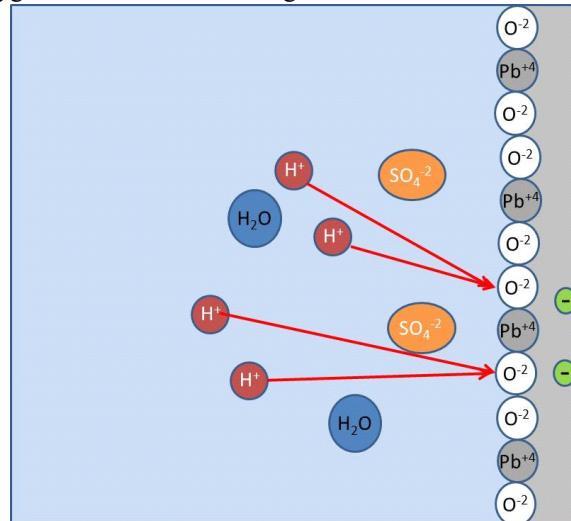
The overall reaction is



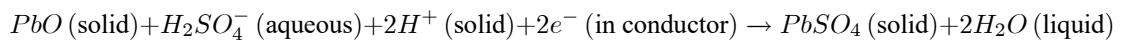
producing lead sulfate on the electrode, some hydrogen ions in solution and some extra electrons that are left in the metal plate.

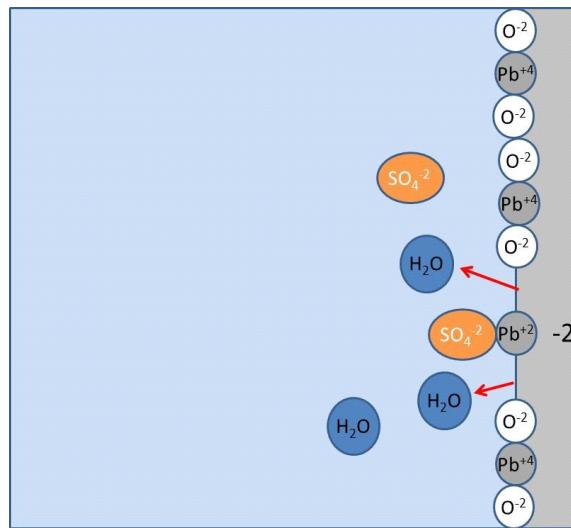


The coated plate's lead dioxide also reacts with the SO_4^{2-} ions and uses the hydrogen ions and the oxygen from the PbO_2 coating.

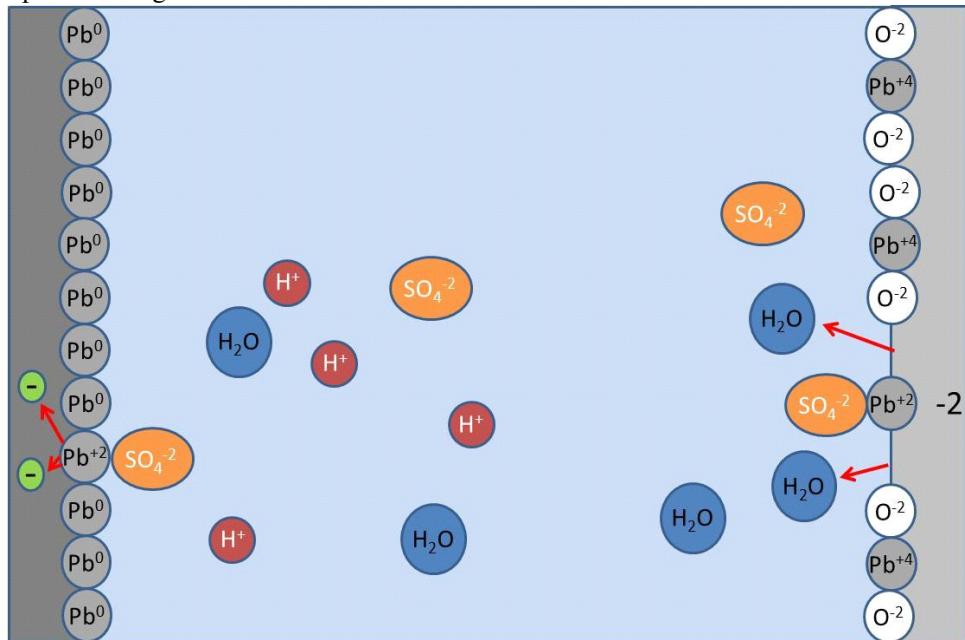


It also uses some electrons from the lead plate. The PbO_2 splits apart and the Pb^{+4} combines with the SO_4^{2-} and the two electrons. The left over O_2 combines with the hydrogens to form water. The reaction equation is





So one lead plate has two extra electrons, and one lacks two electrons. We have separated charge!



If we connect a wire between the plates, the extra electrons from one plate will move to the other plate, and we have formed a current (something we will discuss in detail later). Lead acid batteries are rechargeable. The recharging process places an electric potential across the two lead plates, and this drives the two chemical reactions backwards.

Now that we see that we can use chemistry to separate charge, let's think about what

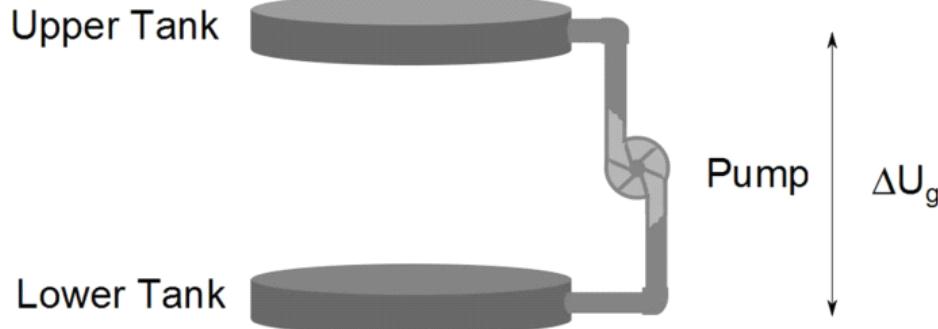
this means for an electric circuit.

$$W_{chem} = \Delta U$$

That work is equivalent to an amount of potential energy, so we have a voltage. That voltage due to the separated charge is

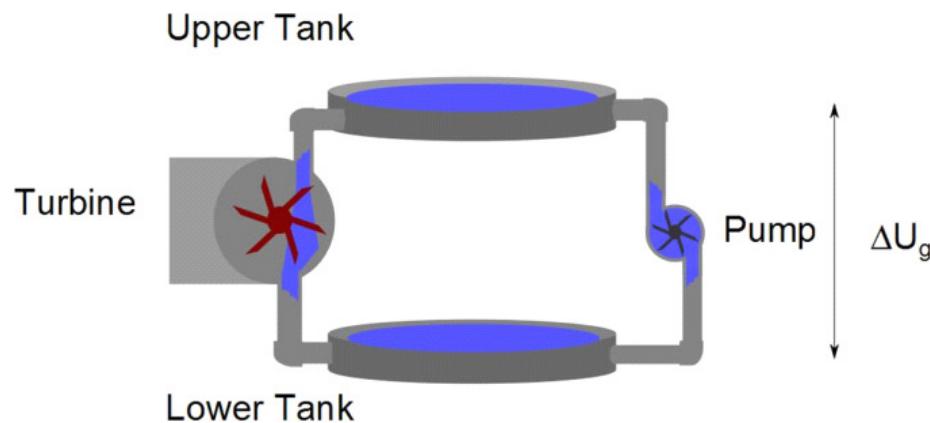
$$\Delta V = \frac{W_{chem}}{q}$$

This is not a chemistry class, so we won't memorize the chemical process that does this. Instead, I would like to give a mechanical analogy.



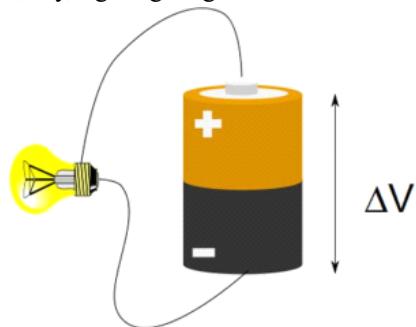
If we have water in a tank and we attach a pump to the tank, we can pump the water to a higher tank. The water would gain potential energy. This is essentially what a battery does for charge. A battery is sort of a “charge pump” that takes charge from a low potential to a high potential.

The water in the upper tank can now be put to work. It could, say, run a turbine.

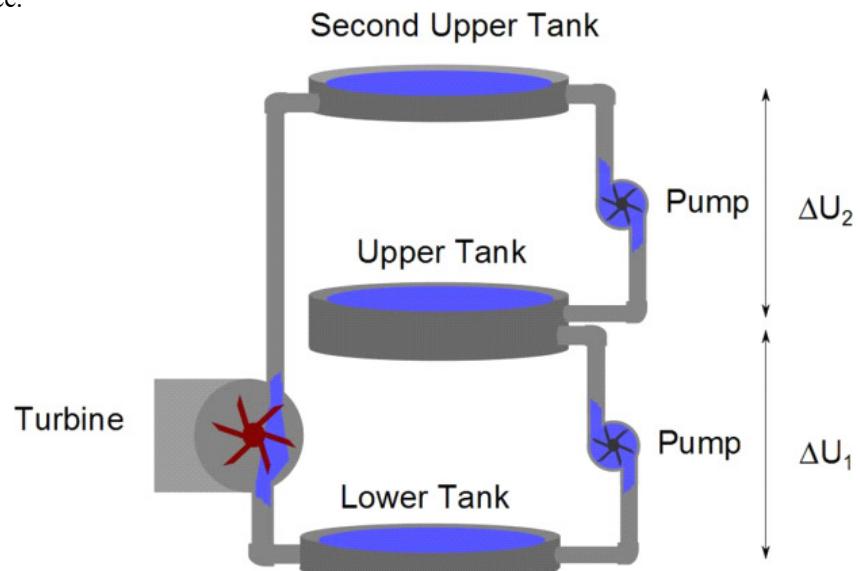


A battery can do the same. The battery “pumps” charge to the higher potential. That

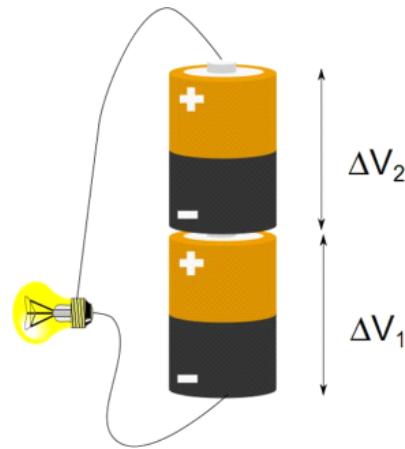
charge can be put to work, say, lighting a light bulb.



Of course, we could string plumps together to gain even more potential energy difference.



likewise we can string two batteries to get a larger electrical potential difference.



If we had more batteries, we would have more potential difference. Each battery “pumping” the charge up to a higher potential. Our analogy is not perfect, but it gives some insight into why stringing batteries together increases the voltage. A television remote likely uses two 1.5 V batteries for a total potential difference from the bottom of the first to the top of the last of

$$\Delta V = 2 \times 1.5 \text{ V} = 3 \text{ V}$$

If you have been introduced to Kirchhoff’s loop law, you may see this as familiar. Kirchhoff said that

$$\Delta V_{loop} = \sum_i \Delta V_i = 0$$

That is, if we go around a loop, we should end up at the same potential where we started. This would be true for our plumbing example. If we start at the lower tank, then travel through the pump to the upper tank, then through the turbine to the lower tank we have

$$\Delta U_{total} = \Delta U_{pump} + \Delta U_{turbine} = 0$$

we are at the same elevation, we lost all the potential energy we gained by being pumped up when we fell back down through the turbine.

Similarly, the battery pumps the charge up an amount ΔV_{bat} and it “falls” down an amount ΔV_{light} returning to where it started

$$\Delta V_{total} = \Delta V_{bat} + \Delta V_{light}$$

This is just conservation of energy. As we go around the loop we must neither create nor destroy energy. We can convert work into potential through the pump or battery, then we can create movement of water or charge and even useful work by letting the

charge or water “fall” back down to the initial state. The change in energy must be zero if there is no loss mechanism. Eventually we must allow some loss to occur, but for now we have ideal batteries and wires and lights, so energy is conserved.

We have a historic name for a charge pump like a battery. We call it an *emf*. This is pronounced “ee em eff,” that is, we say the letters. Emf used to stand for something, but that something has turned out to be a poor model for electric current, but the letters describing a charge pump persist. This is a little like Kentucky Fried Chicken changing its name to KFC because now they bake chicken (and no one wants to think about eating fried foods now days). The letters are the name.

Next lecture we will complete our task. In this lecture we discussed finding the potential if we know the field. Next lecture we will find out how to calculate the field if we know the potential.

Basic Equations

33 Calculating fields from potentials

Fundamental Concepts

- To find the field knowing the potential, we use $\vec{E} = - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) V$
- The gradient shows the direction of steepest change
- The potential of conductors in equilibrium

Finding electric field from the potential

We did part-one of relating fields to potentials in the last lecture. Now it is time for part two, obtaining the electric field from a known potential. Starting with

$$\Delta V = - \int_A^B \vec{E} \cdot d\vec{s}$$

we realize that we should be able to write the integrand as a small bit of potential

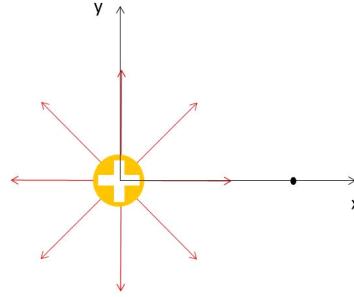
$$\begin{aligned} dV &= -\vec{E} \cdot d\vec{s} \\ &= -E_s ds \end{aligned}$$

where E_s is the component of the electric field in the \hat{s} direction. We can rearrange this

$$E_s = -\frac{dV}{ds}$$

This tells us that the magnitude of our field is the change in electric potential. Of course, \vec{E} is a vector and V is not. So the best we can do is to get the magnitude of the component in the \vec{s} direction.

We can try this out on a geometry we know, say, a point charge along the x -axis



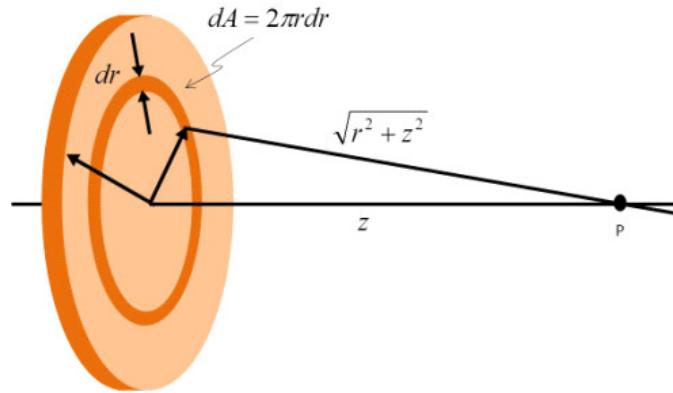
We know the potential will be

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{x}$$

then we can try

$$\begin{aligned} E_s &= -\frac{dV}{dx} = -\frac{d}{dx} \frac{1}{4\pi\epsilon_0} \frac{q}{x} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{x^2} \end{aligned}$$

which gives us just what we expected!



Let's try another. Let's find the electric field due to a disk of charge along the axis. We have done this problem before. We know the field should be

$$E_z = \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{z}{\sqrt{a^2 + z^2}} \right) \quad (33.1)$$

and in the previous lectures we found the potential to be

$$V = \frac{\eta}{2\epsilon_0} \left(\sqrt{a^2 + z^2} - z \right) \quad (33.2)$$

Now can we find the electric field at P from V ? Let's start by finding the z -component

of the field, E_z

$$E_z = -\frac{dV}{dz} \quad (33.3)$$

$$= -\frac{d}{dz} \left(\frac{\eta}{2\epsilon_o} \left(\sqrt{a^2 + z^2} - z \right) \right) \quad (33.4)$$

$$= -\frac{d}{dz} \frac{\eta}{2\epsilon_o} \sqrt{a^2 + z^2} + \frac{d}{dz} \frac{\eta}{2\epsilon_o} z \quad (33.5)$$

$$= -\frac{\eta}{2\epsilon_o} \frac{d}{dz} \sqrt{a^2 + z^2} + \frac{\eta}{2\epsilon_o} \quad (33.6)$$

$$= -\frac{\eta}{2\epsilon_o} \frac{z}{\sqrt{a^2 + z^2}} + \frac{\eta}{2\epsilon_o} \quad (33.7)$$

$$E_z = \frac{\eta}{2\epsilon_o} \left(1 - \frac{z}{\sqrt{a^2 + z^2}} \right) \quad (33.8)$$

or

$$E_z = \frac{2\pi\eta}{4\pi\epsilon_o} \left(1 - \frac{z}{\sqrt{a^2 + z^2}} \right) \quad (33.9)$$

But remember that this situation is highly symmetric. We can see by inspection that all the x and y components will all cancel out. So this is our field! And it is just what we found before.

We can graph these functions to compare them (what would you expect?). To do this we really need values, but instead, let's play a clever trick that some of you will see in advanced or older books. I am going to substitute in place of z the variable $u = \frac{z}{a}$.

Then

$$\begin{aligned} V &= \frac{\eta}{2\epsilon_{o,e}} \left(\sqrt{a^2 + z^2} - z \right) \\ &= \frac{\eta a}{2\epsilon_o} \left(\sqrt{1 + \frac{z^2}{a^2}} - \frac{z}{a} \right) \\ &= \frac{\eta a}{2\epsilon_o} \left(\sqrt{1 + u^2} - u \right) \end{aligned}$$

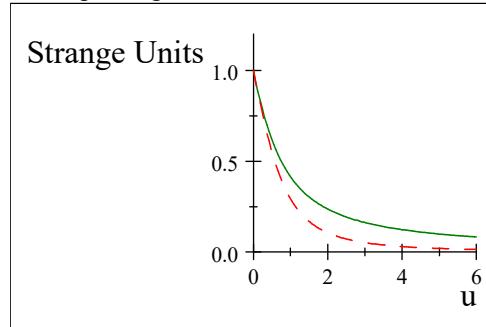
and

$$\begin{aligned}
 E_z &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{z}{\sqrt{a^2 + z^2}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{z}{a\sqrt{1 + \frac{z^2}{a^2}}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{z}{a\sqrt{1 + \frac{z^2}{a^2}}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{\frac{z}{a}}{\sqrt{1 + \frac{z^2}{a^2}}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{u}{\sqrt{1 + u^2}} \right)
 \end{aligned} \tag{33.10}$$

Both my equation for V and for E_z now are in the form of a set of constants times a function of u .

$$\begin{aligned}
 V &= \frac{\eta a}{2\epsilon_0} \left(\sqrt{1 + u^2} - u \right) \\
 &= \frac{\eta a}{2\epsilon_0} f(u) \\
 E_z &= \frac{2\pi\eta}{4\pi\epsilon_0} \left(1 - \frac{u}{\sqrt{1 + u^2}} \right) \\
 &= \frac{2\pi\eta}{4\pi\epsilon_0} g(u)
 \end{aligned} \tag{33.11}$$

If I plot V in units of $\frac{\eta a}{2\epsilon_0}$ (the constants out in front) I can see the shape of the curve. It is the function of $f(u)$. I can compare this to E_z in units of $\frac{2\pi\eta}{4\pi\epsilon_0}$. The shape of E_z will be $g(u)$. Of course we are plotting terms of u .



Now we can ask, is this reasonable? Does it look like the E -field (red dashed line) is the right shape for the derivative of the potential (solid green line)? It is also comforting to see that as u (a function of z) gets larger the field falls off to zero and so does the potential as we would expect. When V (green solid curve) has a large slope, E_z is a

large number (positive because of the negative sign in the equation

$$E_s = -\frac{dV}{ds}$$

and when V is fairly flat, E_z is nearly zero. Our strategy for finding E from V seems to work.

Geometry of field and potential

You should probably worry that so far our equation

$$E_s = -\frac{dV}{ds}$$

is only one dimensional. We know the electric field is a three dimensional vector field.

We may find situations where we need two or three dimensions. But this is easy to fix.

Our equation

$$E_s = -\frac{dV}{ds}$$

gives us the field magnitude along the \hat{s} direction. Let's choose this to be the \hat{x} direction. Then

$$E_x = -\frac{dV}{dx}$$

is the x -component of the electric field. Likewise

$$E_y = -\frac{dV}{dy}$$

$$E_z = -\frac{dV}{dz}$$

The total field will be the vector sum of it's components

$$\begin{aligned}\vec{E} &= E_x \hat{i} + E_y \hat{j} + E_z \hat{k} \\ &= -\frac{dV}{dx} \hat{i} - \frac{dV}{dy} \hat{j} - \frac{dV}{dz} \hat{k}\end{aligned}$$

Question 223.33.1 which we can cryptically write as

Question 223.33.2

$$\vec{E} = - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) V$$

The odd group of operations in the parenthesis is call a *gradient* and is written as

$$\vec{\nabla} = \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right)$$

using this we have

$$\vec{E} = -\vec{\nabla}V$$

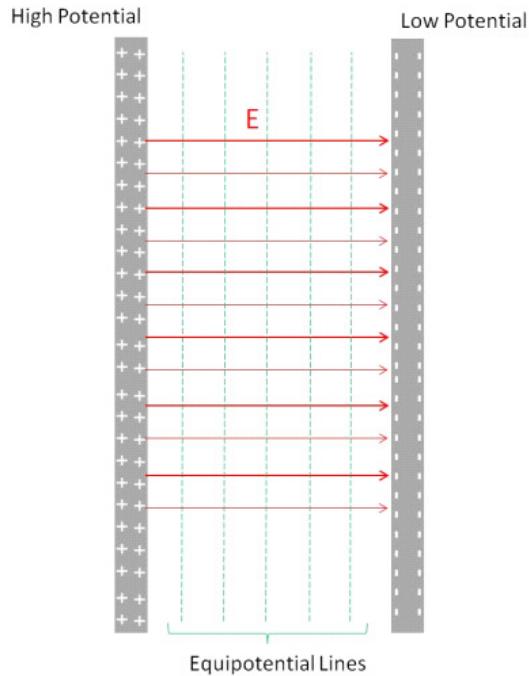
which is how the relationship is stated in higher level electrodynamics books. But what does it mean?

The gradient is really kind of what it sounds like. If you go down a steep grade, you will notice you are going down hill and will notice if you are going down the steepest

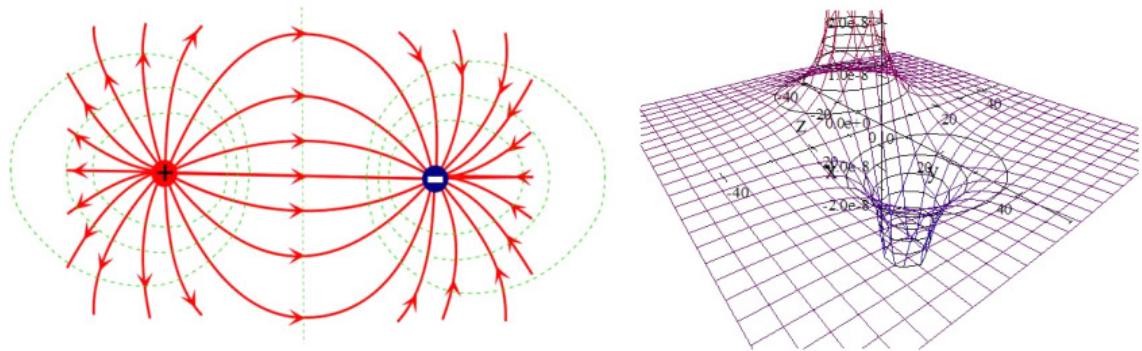
Stamp in a circle: mimic a blindfolded person swiveling on one foot and testing the slope with the other

part of the hill. The gradient finds the direction of steepest decent. That is, the direction where the potential changes fastest. This is like looking from the top of the hill and taking the steepest way down! Our relationship tells us that the electric field points in this steepest direction, and the minus sign tells us that the electric field points down hill away from a positive charge, never up hill (think of the acceleration due to gravity being negative). Let's see if this makes sense for our geometries that we know.

Here is our capacitor. We see that indeed the field points from the high potential to the low potential. The steepest way "down the hill" is perpendicular to the equipotential lines.



We also know the shape of the field for a dipole. The equipotential lines we have seen before.



But now we can see that the field lines and equipotential lines are always perpendicular and the field points “down hill.” The meeting of the field and equipotential lines at right angles is not a surprise. Think again about our mountain



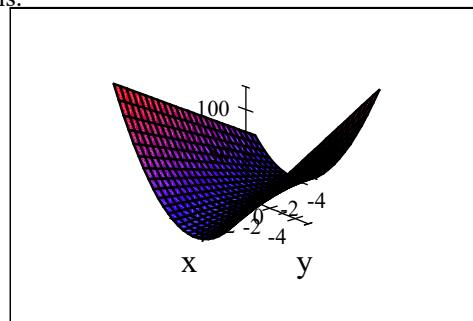
Map courtesy USGS, Picture is in the Public Domain.

The steepest path is always perpendicular to lines of equal potential energy.

We should try another example of finding the field from the gradient. Suppose we have a potential that varies as

$$V = 3x^2 + 2xy$$

I don’t know what is making this potential, but let’s suppose we have such a potential. It would look like this.



what is the electric field?

$$\vec{E} = -\vec{\nabla}V$$

or

$$\vec{E} = - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) V$$

so

$$\vec{E} = - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) (3x^2 + 2xy)$$

$$\begin{aligned}\vec{E} &= - \left(\hat{i} \frac{d}{dx} (3x^2 + 2xy) + \hat{j} \frac{d}{dy} (3x^2 + 2xy) + \hat{k} \frac{d}{dz} (3x^2 + 2xy) \right) \\ &= - \left(\hat{i} (6x + 2y) + \hat{j} \frac{d}{dy} (2xy) + 0 \right)\end{aligned}$$

This example shows how to perform the operation, but it does not give much insight. We have learned to work with our standard charge configurations, and this is really not one of them. So we don't have much intuitive feel for this electric field that we found.

To gain more insight, let's return to finding the point charge field from the point charge potential. The potential for a point charge is

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{r}$$

And of course we know that the field is

$$E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \hat{r}$$

but we want to show this using

$$\vec{E} = -\vec{\nabla}V$$

So

$$\begin{aligned}
 \vec{\mathbf{E}} &= - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) V \\
 &= - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) \frac{1}{4\pi\epsilon_o} \frac{Q}{r} \\
 &= - \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) \frac{1}{4\pi\epsilon_o} \frac{Q}{\sqrt{x^2 + y^2 + z^2}} \\
 &= - \frac{Q}{4\pi\epsilon_o} \left(\frac{d}{dx} \hat{i} + \frac{d}{dy} \hat{j} + \frac{d}{dz} \hat{k} \right) \frac{1}{\sqrt{x^2 + y^2 + z^2}} \\
 &= - \frac{Q}{4\pi\epsilon_o} \left(-\frac{x}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \hat{i} - \frac{y}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \hat{j} - \frac{z}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \hat{k} \right) \\
 &= \frac{Q}{4\pi\epsilon_o} \frac{(x\hat{i} + y\hat{j} + z\hat{k})}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \\
 &= \frac{Q}{4\pi\epsilon_o} \frac{(x\hat{i} + y\hat{j} + z\hat{k})}{(x^2 + y^2 + z^2) \sqrt{(x^2 + y^2 + z^2)}} \\
 &= \frac{1}{4\pi\epsilon_o} \frac{Q}{r^2} \frac{(x\hat{i} + y\hat{j} + z\hat{k})}{\sqrt{(x^2 + y^2 + z^2)}} \\
 &= \frac{1}{4\pi\epsilon_o} \frac{Q}{r^2} \hat{\mathbf{r}}
 \end{aligned}$$

but really, this is a bit of a mess, we don't want to do such a problem in rectangular coordinates. We could write ∇ in spherical coordinates (something we won't derive here, but you should have seen in M215 or M316).

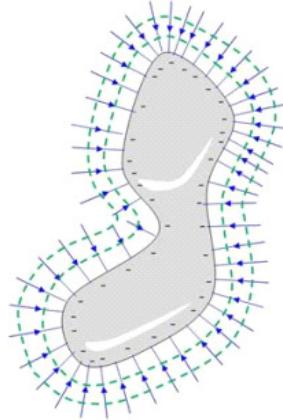
$$\vec{\nabla} = \hat{\mathbf{r}} \frac{\partial}{\partial r} + \hat{\theta} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi}$$

Let's try this out on our point charge potential. We have

$$\begin{aligned}
 \vec{\mathbf{E}} &= - \left(\hat{\mathbf{r}} \frac{\partial}{\partial r} + \hat{\theta} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) V \\
 &= - \left(\hat{\mathbf{r}} \frac{\partial}{\partial r} + \hat{\theta} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \right) \frac{1}{4\pi\epsilon_o} \frac{Q}{r} \\
 &= - \frac{Q}{4\pi\epsilon_o} \left(-\frac{1}{r^2} \hat{\mathbf{r}} + 0 + 0 \right) \\
 &= \frac{1}{4\pi\epsilon_o} \frac{Q}{r^2} \hat{\mathbf{r}}
 \end{aligned}$$

just as we expected. But this time the math was much easier. If we can, it is a good idea to match our expression for $\vec{\nabla}$ to the geometry of the system. A good vector calculus book or a compendium of math functions will have various versions of $\vec{\nabla}$ listed.

Conductors in equilibrium again



Question 223.33.3

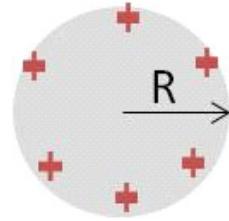
We know that there is no field inside a conductor in electrostatic equilibrium, but we should ask what that means for the electric potential. To build circuits or electronic actuators, we will need to know this. Let's start again with

$$\Delta V = - \int_A^B \mathbf{E} \cdot d\mathbf{s} \quad (33.12)$$

and since the field $E = 0$ inside the conductor, then inside

$$\Delta V_{inside} = 0 \quad (33.13)$$

On the surface we see that there is a potential, since there is a field. If we take our spherical case,



and observe the potential as we go away from the center, we expect the potential to be constant up to the surface. Then as we reach the surface, we know from Gauss' law that the field will be

$$E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2}$$

like a point charge, so the potential at the surface must be

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{R}$$

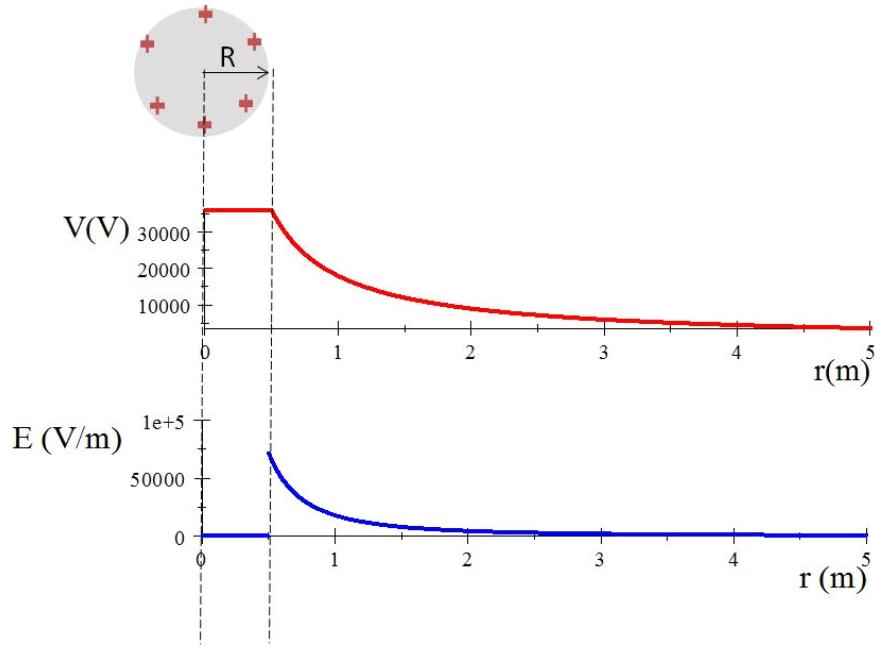
where $r = R$, the radius of our sphere. As we move into the sphere from the surface, the potential must not change. The interior will have the potential

$$V_{inside} = \frac{1}{4\pi\epsilon_0} \frac{Q}{R} \quad (33.14)$$

Outside, of course, the potential will drop like the potential due to a point charge. We expect

$$V = \frac{1}{4\pi\epsilon_0} \frac{Q}{r} \quad (33.15)$$

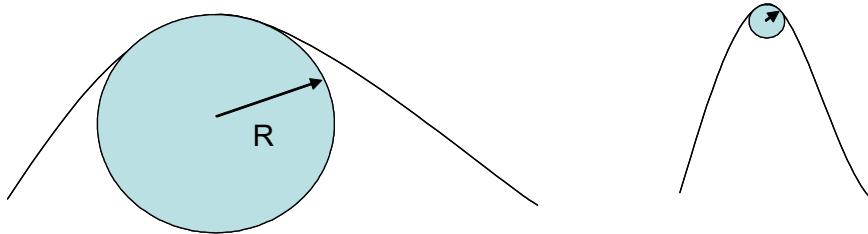
For a sphere of radius $R = 0.5$ m carrying a charge of 0.000002 C (about what our van de Graaff holds) we would have the situation graphed in the following figure:



This is an important point. For a conductor, the electric potential everywhere inside the conductive material is exactly the same once we reach equilibrium. This is just what we want for capacitors or electrodes or electrical contacts in circuits.

Non spherical conductors

The field is stronger where the field lines are closer together. One way to describe this is to use a radii of curvature. That is, suppose we try to fit a small circle into a bump on the surface of a conductor.



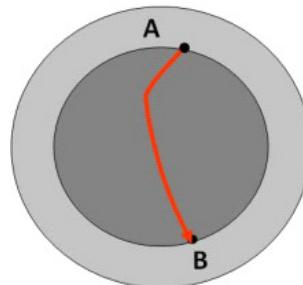
In the figure there are two bumps shown with circles fit into them. The bump on the right has a much smaller radius circle than the one on the left. The radius of the circle that fits into the bump is the radius of curvature of the bump. From what we have said, the bump on the right will have a much stronger field strength near it than the bump on the left.

Where there is a lot of charge on a conductor, and the field is very high, electrons from random ionizations of air molecules near the conductor are accelerated away from the conductor. These electrons hit other atoms, ionizing them as well. We get a small avalanche of electrons. Eventually the electrons recombine with ionized atoms, producing an eerie glow. This is called *corona discharge*. It can be used to find faults in high tension wires and other high voltage situations.

Coronal Discharge Clips

Cavities in conductors

Suppose we have a hollow conductor with no charges in the cavity. What is the field? We know from using Gauss' law what the answer should be, but let's do this using potentials.



All the parts of the conductor will be at the same potential. So let's take two points, *A* and *B*, and compute

$$V_A - V_B = - \int_A^B \mathbf{E} \cdot d\mathbf{s}$$

We know that $V_A - V_B = 0$ because V_A must be the same as V_B . So for every path, *s*,

we must have

$$-\int_A^B \mathbf{E} \cdot d\mathbf{s} = \mathbf{0}$$

We can easily conclude that E must equal zero.

So as long as there are no charges inside the cavity, the cavity is a net field free zone.

It is often much easier to find the potential, and from the potential, find the field. Much of the study of electrodynamics uses this approach. This is because it is more straight-forward to differentiate than it is to integrate. Some of you may use massive computational programs to predict electric fields. They often use differential equations in the potential to find the field rather than integral equations to find the field directly.

Basic Equations

34 Capacitance

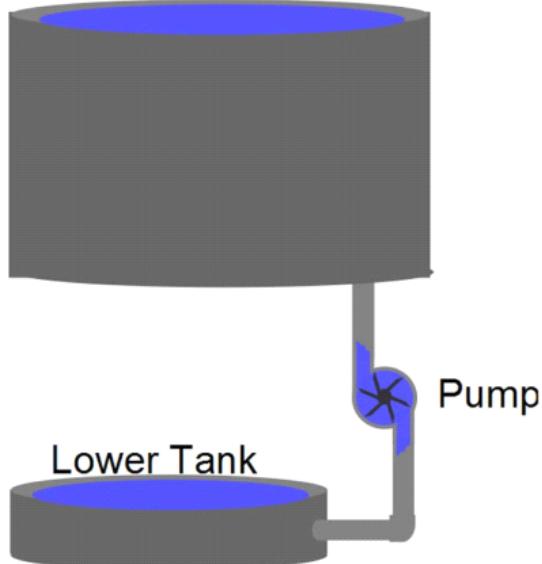
Fundamental Concepts

- The charge on a capacitor is proportional to the potential difference $Q = C\Delta V$
- The constant of proportionality is called the capacitance and for a parallel plate capacitor, it is given by $C = \frac{A}{d}\epsilon_0$
- In parallel capacitors capacitances add $C_{eq} = C_1 + C_2$
- In series capacitors capacitances combine as $\frac{1}{C_{tot}} = \frac{1}{C_1} + \frac{1}{C_2}$

Capacitance and capacitors

Consider the following design for a pump-tank system.

Upper Tank



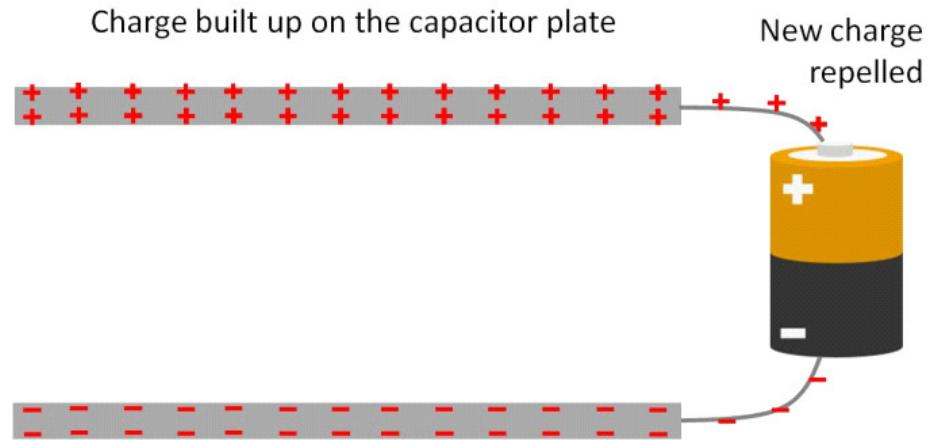
This may not be an optimal design. At first there is no problem, water flows into the upper tank just fine. But once the upper tank begins to fill, the water already in the

upper tank will make it harder to pump in more water. As the tank fills, the pressure at the bottom increases, and it takes more work for the pump to overcome the increasing pressure.

Something analogous happens when a capacitor is connected to a battery.



At first the charge is free to flow to the plates, but as the charge builds, it takes more work to bring on successive charges.



The charges repel each other, so the charge already on a capacitor plate repels the new charge arriving from the battery. The repelling force gets larger until finally the force repelling the charge balances the force driving the charge from the battery and the charge stops flowing onto the capacitor.

A capacitor is made from two plates. For us, let's assume they are semi-infinite sheets of charge. Of course this is not exactly true, but it is not too wrong near the center of the plates. And we know quite a lot about semi-infinite sheets of charge because they are one of our standard charge configurations. We know the field for each sheet is

$$E = \frac{\eta}{2\epsilon_0}$$

and that for two sheets, one with $+\eta$ and one with $-\eta$ the field in between will be

$$E = \frac{\eta}{\epsilon_o}$$

We also know the potential difference between the two plates is just

$$\Delta V = Ed$$

where E is our electric field and d is the capacitor spacing.

We can guess that we will build up charge until the potential energy difference of the capacitor is equal to the potential energy difference of the battery

$$\Delta V_{capacitor} = \Delta V_{battery}$$

because at that point the forces causing the potential energy will be equal.

We can write our electric field between the two plates as

$$E = \frac{\eta}{\epsilon_o} = \frac{Q}{A\epsilon_o}$$

so

$$\Delta V = \frac{d}{A\epsilon_o} Q$$

Then the potential difference is directly proportional to the charge. I want to switch this around, and solve for the amount of charge.

$$Q = \left(\frac{A\epsilon_o}{d} \right) \Delta V$$

Since all the terms in the parenthesis are constants, we could replace them with a constant, C .

$$Q = C\Delta V \quad (34.1)$$

where

$$C = \frac{A}{d}\epsilon_o \quad (34.2)$$

is a constant that depends on the geometry and construction of the plates. This equation tells us that if we build two different sets of plates, say, one circular and one triangular, and we give them the same potential difference (say, connect them both to 12 V batteries) then, if both have the same construction constant C , they will carry the same charge even though their size and shape are different. We can reduce the burden of calculation of how much charge a capacitor can hold but asking the person who manufactured it to calculate the construction constant and mark the value on the outside of the capacitor. Different capacitors may be constructed differently (different A or d values) but so long as the construction constant, C , is the same, the charge amount for a given voltage will be the same.

Question 223.34.1

The electronics field gives this construction constant a name, *capacitance*.

$$C = \frac{Q}{\Delta V} \quad (34.3)$$

The capacitance will have units of C/V but we give this a name all its own, the *Farad* (F). A Farad is a very large capacitance. Many capacitors in electronic devices are measured in microfarads.

Question 223.34.2

Question 223.34.3

Question 223.34.4

Capacitors and sources of potential

Consider what happens when we connect our two parallel plates to the terminals of a battery. Assuming the plates are initially uncharged, charge flows from the battery through the conducting wires and onto the plates. Recall that for a metal, the entire surface will be at the same potential under electrostatic conditions. The charge carriers supplied by the battery will try to achieve electrostatic equilibrium, so we expect the plate that is connected to the positive terminal of the battery to eventually be at the same potential as the positive battery terminal. Likewise for the negative terminal and the plate connected to it.

We can even use our capacitor as a source of electrical power. A camera flash uses capacitors to make the burst of light that illuminates the subject of your photo.



Camera flash unit (Public Domain image by Julo)

Single conductor capacitance

Physicists can't leave a good thing alone. We often calculate the capacitance of a single conductor! If the geometry is simple we can easily do this. It is not immediately obvious that a single conductor should even have a capacitance, so it might be a problem if you forgot this in a design problem for an unusual device.

As an example, let's take a sphere. We will assume there is a spherical conduction shell that is infinitely far away. This configuration gives exactly the same field lines that the

charged sphere gives on its own, but the mental picture is helpful. The imaginary shell will give $V = 0$ (we set our zero potential at $r = \infty$). The potential of the little sphere we know must be just like the potential of a point charge if we are outside of the sphere

$$V = k_e \frac{Q}{r}$$

for $r = R$, the radius of our little sphere. Then

$$\Delta V = k_e \frac{Q}{R} - 0 = k_e \frac{Q}{R}$$

so

$$C = \frac{Q}{\Delta V} = \frac{Q}{k_e \frac{Q}{R}} = \frac{R}{k_e} = 4\pi\epsilon_o R \quad (34.4)$$

This is the capacitance of a single sphere. Note that C only depends on geometry! not on Q , just as we would expect.

But why would we care? This says that even if we just connect a ball to, say, the positive terminal of a battery, that there will be some capacitance. This capacitance will limit the flow of charge to the ball. So it will take time to charge even a single conductor. This is always true when a device is initially connected to a power source. Often we can ignore such “transient” effects because the charging times are still small. But in special cases, this may not be possible because the changing voltage or charge could damage sensitive equipment. So although this is rarely a problem, it is good to keep in the back of our minds.

Capacitance of two parallel plates

The capacitance of single conductors is profound, but more useful to us in understanding common electronic components is the parallel plate capacitor. We found that for parallel plates we also had only geometry factors in the capacitance. Of course, there are other shapes possible. Let's see if we can reason out how the capacitance depends on the geometry.

Plate area

Since the charge will tend to separate to the surface of a conductor, we might expect that if the surface area increases, the amount of charge that the capacitor can hold might increase as well. We see this in our equation for the parallel plate capacitor.

$$C = \frac{A}{d}\epsilon_o$$

Plate separation

We also see that it matters how far apart the plates are placed. The greater the distance, the less the capacitance. This makes some sense. If the plates are farther apart, the Coulomb force is weaker, and less charge can be held in the capacitor, because the force attracting the charges (the force between the charges on the opposite plates) is weaker.

Capacitance of a cylindrical capacitor

We should try some harder geometries. A cylindrical capacitor is a good case to start with



(you will do a sphere in the homework problems). We want to find the capacitance of the cylindrical capacitor. Our strategy will be to find the voltage difference for the capacitor and the amount of charge on the capacitor, and then divide to find C .

$$C = \frac{Q}{\Delta V}$$

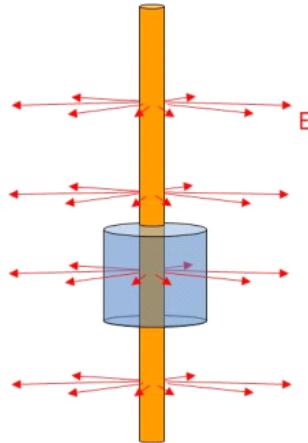
Let's begin with our equation relating potential change to field.

$$V_b - V_a = - \int_a^b \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{s}} \quad (34.5)$$

Let's assume that there is a linear charge density, λ , along the cylinder with the center positive and the outside negative. Then

$$\Phi_E = \oint \tilde{\mathbf{E}} \cdot d\tilde{\mathbf{A}} \quad (34.6)$$

where I will choose a Gaussian surface that is cylindrical around the central conductor.



This is nice, since the field will be radially out from the conductor (ignoring the end effects) and so no field will pass through the end caps of the Gaussian surface ($\mathbf{E} \cdot d\mathbf{A} = 0$ on the end caps). Moreover, the field strikes the surface at right angles ($\mathbf{E} \cdot d\mathbf{A} = EdA$ on the side of the cylinder), and will have the same magnitude all the way around so

$$\begin{aligned}\Phi_E &= E \oint dA \\ &= EA\end{aligned}$$

Now we know from Gauss' law that

$$\Phi_E = \frac{Q_{in}}{\epsilon_0}$$

where

$$Q_{in} = \lambda h$$

and where h is the height of our Gaussian surface, so

$$\begin{aligned}\Phi_E &= \frac{\lambda h}{\epsilon_0} = E 2\pi r h \\ \frac{\lambda}{2\pi r \epsilon_0} &= E\end{aligned}$$

Now, knowing our field, and taking a radial path from a to b , we can take

$$\begin{aligned}V_b - V_a &= - \int_a^b \frac{\lambda}{2\pi r \epsilon_0} dr \\ &= -\frac{\lambda}{2\pi \epsilon_0} \int_a^b \frac{1}{r} dr \\ &= -\frac{\lambda}{2\pi \epsilon_0} \ln\left(\frac{b}{a}\right)\end{aligned}$$

Using this, we can find the capacitance. We have a negative value for ΔV , but this is just due to our choice of making the center of the concentric cylinders positive and the outside negative. We chose the zero point on the positive center. The amount of

potential change going from a to b is just $|\Delta V|$. Then in finding the capacitance using

$$Q = C\Delta V$$

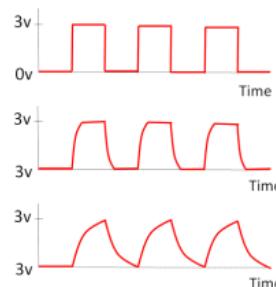
We want just the value of ΔV so we will plug in the absolute value of our result.

$$|\Delta V| = \frac{\lambda}{2\pi\epsilon_0} \ln\left(\frac{b}{a}\right)$$

Then, solving for C gives

$$\begin{aligned} C &= \frac{Q}{\Delta V} \\ &= \frac{Q}{\frac{\lambda}{2\pi\epsilon_0} \ln\left(\frac{b}{a}\right)} \\ &= \frac{Q}{\frac{Q}{2\pi h\epsilon_0} \ln\left(\frac{b}{a}\right)} \\ &= \frac{2\pi h\epsilon_0}{\ln\left(\frac{b}{a}\right)} \end{aligned}$$

Wow! That was fun! But more importantly, this is a coaxial cable geometry, and we can see that coaxial cable will have some capacitance and that that capacitance will depend on the geometry of the cable including its length and width. This capacitance can affect signals sent through the cable. Later in our course we will see why. But for now just know that if I combine a resistor and a capacitor together it takes more time for the charge to move. So in our signal cable the signal will get distorted.



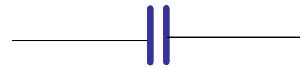
Increasing amounts of distortion in a signal due to increasing cable capacitance.

The nice square pulses that represent digital data will be distorted, and in extreme cases, undetectable. When designing data lines, this capacitance of the cable must be taken into account.

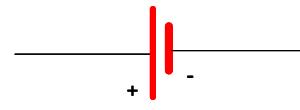
Combinations of Capacitors

Question 223.34.5

We don't want to have to do long calculations to combine capacitors that we buy from an electronics store. It would be convenient to come up with a way to combine capacitors using a simple rule.



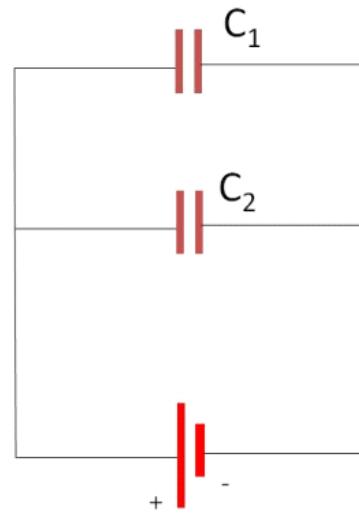
Capacitor



Battery

Figure 34.19.

We need a simple way to write capacitors in our homework problem drawings, here are the usual symbols for capacitor and battery. Using these symbols, let's consider two capacitors as shown below.



Remember that a conductor will be at the same potential over all of its surface. If we connect the capacitors as shown then all of the left half of this diagram will be at the positive potential of the battery terminal. Likewise, the right side will all be at the same potential. It is like we increased the area of the capacitor C_1 buy adding in the area of capacitor C_2 .

$$C = \frac{A_1 + A_2}{d} \epsilon_o = \frac{A_1}{d} \epsilon_o + \frac{A_2}{d} \epsilon_o$$

So we may write a combined capacitance for this set up of

$$C_{eq} = C_1 + C_2 \quad (34.7)$$

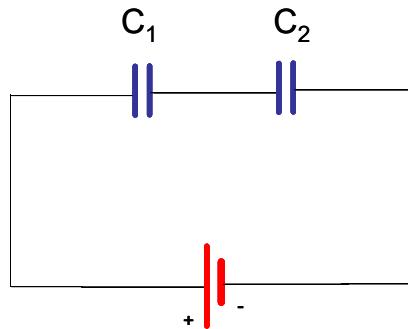


Figure 34.20.

We call this set up a *parallel* circuit. This means that each of the capacitors are hooked directly to the terminals of the battery.

But suppose we hook up the capacitors as in the next drawing. Now we expect the left hand side of C_1 to be at the positive potential of the positive terminal of the battery. We expect the right side of C_2 to be at the same potential as the negative side of the battery. What happens in the middle?

We can see that we will have negative charge on the right hand plate of C_2 and positive charge on the left plate of C_1 . This must cause there to be a positive charge on the right plate of C_1 and a negative charge on the left plate of C_2 . Moreover, all the charges will have the same magnitude. That means each of the plates will have a potential difference

$$\Delta V_1 = \frac{Q}{C_1}$$

and

$$\Delta V_2 = \frac{Q}{C_1}$$

But the total potential difference is ΔV of the battery, then

$$\Delta V = \Delta V_1 + \Delta V_2$$

We can again define an equivalent capacitance.

$$\Delta V = \frac{Q}{C_{tot}}$$

then

$$\begin{aligned} \Delta V &= \Delta V_1 + \Delta V_2 \\ \frac{Q}{C_{tot}} &= \frac{Q}{C_1} + \frac{Q}{C_2} \end{aligned}$$

The Q s are all the same. So

$$\frac{1}{C_{tot}} = \frac{1}{C_1} + \frac{1}{C_2} \quad (34.8)$$

We call this type of set up a *series circuit* because the capacitors came one after the other as you go from one side of the battery to the other.

Now after all this you might ask yourself how to know the capacitance of the parts you buy to build things. They are designed by engineers and tested at the factory, and the capacitance is usually printed on the side of the device. You can, of course, devise a test circuit based on what we have learned that could test the capacitance.

Basic Equations

35 Dielectrics and Current

Fundamental Concepts

- Dielectrics and capacitors
- Microscopic nature of electric current
- Current direction is defined as the direction positive charges would go, regardless of the actual sign of the charge.
- In a capacitor, the stored energy is $W = \frac{1}{2}C\Delta V^2$
- The energy density in the electric field is $u = \frac{1}{2}\epsilon_o E^2$

Energy stored in a capacitor

We have convinced ourselves that ΔV is the change in potential energy per unit charge, so when a capacitor is charged, and the wires connecting it to the battery are removed, is there potential energy “stored” in the capacitor? The answer is yes, and we can see it by considering what would happen if we connected a wire (no battery) between the two plates. Charge would rush from one plate to the other. This is like storing a tank of water on a hill. If we connect a pipe from the tank at the top of the hill to a tank at the bottom of the hill, the water will rush through the pipe to the lower tank.

BE CAREFUL, you are enough of a conductor that by touching different ends of a capacitor you could create a serious current through your body. The capacitors in old computer monitors or old TV sets can store enough charge to kill you!

But how much energy would there be stored in the capacitor? Clearly it must be related to the amount of energy it takes to move the charge onto the plates. By analogy, the energy stored in the water was the minimum amount of energy it took to pump the water to the upper tank (mgh). It is the minimum, because our pipes might have some resistance, and then we would have to include more work to overcome the resistance.

But for a capacitor it is a little bit more tricky. When the capacitor is not charged, it

takes no work (or very little) to move charge from one plate to the other. But once there is a charge there is an electric field between the plates (think of my poorly designed water storage system from the beginning of last lecture). This creates a potential difference. And we must fight against this potential difference to add more charge. This is sort of like transferring rocks up a hill. The more rocks that we carry, the higher the hill gets, and the more work it takes to bring up more rocks.

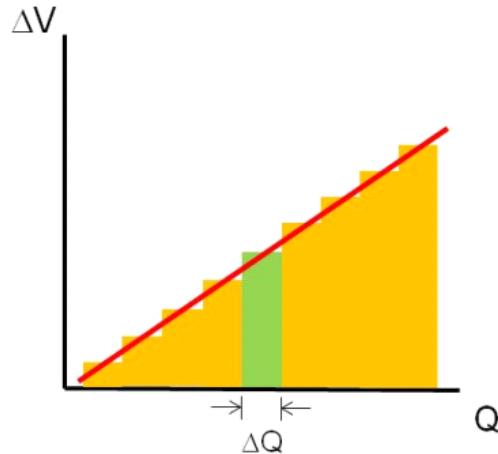
From our formula

$$W = q(V_B - V_A)$$

we can see that if we have just a small amount of charge, ΔQ , we will have a small amount of work

$$\Delta W = \Delta Q \Delta V$$

to move it onto the capacitor.²⁸ If we start with no charge, then go in small ΔQ steps, we would see a potential rise as shown in the graph below.



The quantity $\Delta Q \Delta V$ is the area of the shaded (green) rectangle. So ΔW is given by the area of a rectangle under a stair-step on our graph. The shaded rectangle is just one of many rectangles in the graph. we can write

$$C = \frac{\Delta Q}{\Delta V} \quad (35.1)$$

or

$$\Delta V = \frac{\Delta Q}{C}$$

As ΔQ gets small we can go to a continuous charge model

$$\Delta W = \Delta Q \Delta V$$

²⁸ Agh! here ΔQ is a small amount of charge, and ΔV is $V_f - V_i$. We have used Δ in two different ways in the same equation.

We can replace the small unit of charge ΔQ with a continuous variable q to obtain

$$dW = dq (\Delta V)$$

Recall that

$$\Delta V = \frac{q}{C}$$

so we can write dW as

$$\begin{aligned} dW &= dq \left(\frac{q}{C} \right) \\ dW &= \frac{1}{C} q dq \end{aligned}$$

Of course, we will integrate this

$$\begin{aligned} W &= \int_0^Q \frac{1}{C} q dq \\ W &= \frac{1}{C} \int_0^Q q dq \\ &= \frac{Q^2}{2C} \end{aligned} \tag{35.2}$$

or sometimes using

$$Q = C\Delta V$$

this is written as

$$W = \frac{1}{2} C \Delta V^2 \tag{35.3}$$

There is a limit to how much energy we can store. That is because even air can conduct charge if the potential difference is high enough. We call this air conduction a spark or coronal discharge. At some point charge jumps from one plate to another through the air in between. If the potential difference is very high, the Coulomb force between the charges on opposite plates will force charge to leave one plate and jump to the other even if there is no air!

Question 223.34.6

Question 223.34.7

Field storage

We usually consider the energy stored in the capacitor to be stored in the electric field. The field is proportional to the amount of charge and related to the potential energy, so this seems reasonable. Let's find the potential energy stored in the field in the capacitor. Recall for an ideal parallel plate capacitor

$$\Delta V = Ed$$

and

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

We assume that energy provided by the work to move the charges on the capacitor is all

stored as potential energy, so

$$U_{stored} = \frac{1}{2}C\Delta V^2 \quad (35.4)$$

then

$$\begin{aligned} U_{stored} &= \frac{1}{2} \left(\epsilon_0 \frac{A}{d} \right) (Ed)^2 \\ &= \frac{1}{2} \epsilon_0 A d E^2 \end{aligned}$$

We often define an energy density

$$u = \frac{U_{stored}}{\nabla}$$

In this case the volume ∇ is just Ad so

$$u = \frac{1}{2} \epsilon_0 E^2 \quad (35.5)$$

This is the density of energy in the electric field. It turns out that this is a general formula (not just true for ideal parallel plate capacitors).

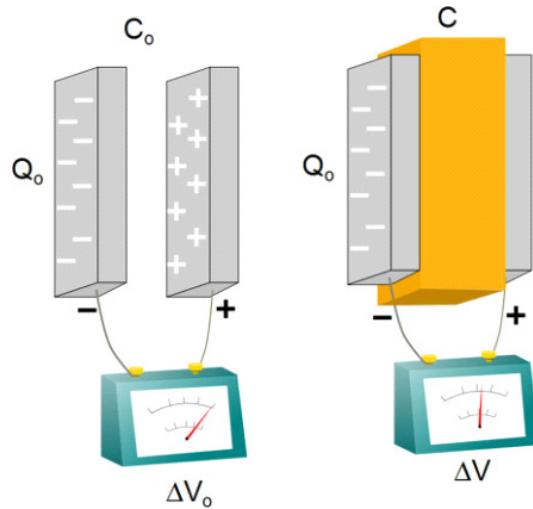
This is a step toward our goal of showing that electric fields are a physically real thing. They can store energy, so they must be a real thing.

Dielectrics and capacitors

Question 223.35.1

We should ask ourselves a question about our capacitors, does it matter that there is air in between the plates? For making capacitors, it might be convenient to coat two sides of a plastic block with metal and solder wires to the coated sides. Does the plastic have an effect?

Plastic is an insulator, and another name for “insulator” is *dielectric*. If we perform the experiment, we will find that when a dielectric is placed in the plates, the potential difference decreases!

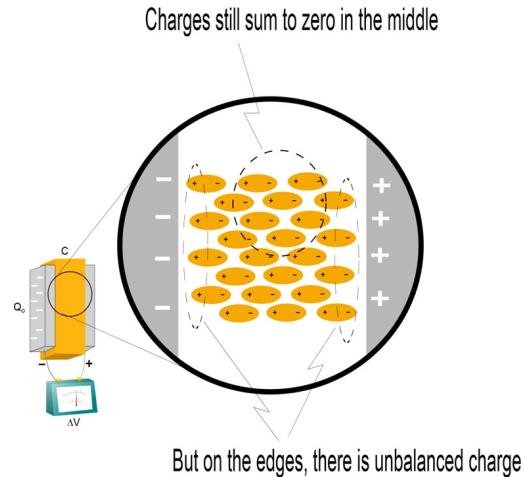


We are lucky, though, from experimentation we have found that it seems to decrease in a nice, linear way. We can write this as

$$\Delta V = \frac{\Delta V_{wo}}{\kappa} \quad (35.6)$$

where κ is a constant that depends on what material we choose as our dielectric²⁹ and ΔV_{wo} is the potential difference without the dielectric (the subscripts *wo* will stand for “without the dielectric.” But what is happening?

The plates of the capacitor are becoming charged. These charges will polarize the material in the middle.



Question 223.35.2

Notice how the polarized molecules or atoms still have a net zero charge in the middle,

²⁹ This symbol κ , is the greek letter “kappa.”

Unbalanced Handedness Demo, Stick out your hands, one side of room has extra left hands, one side extra right hands

Question 223.35.3

but on the ends, there is a net charge. It is like we have oppositely charged plates next to our capacitor plates. That reduces the net charge seen by the capacitor, and so the potential difference is less. There is effectively less separated charge.

But since our capacitor is not connected to a battery or any other electrical device, the amount of actual charge on the capacitor plates can't have changed, so if ΔV changed, but Q did not, then since

$$Q = C\Delta V$$

we suspect the material properties part—the capacitance—must have changed.

$$C = \frac{Q_{wo}}{\Delta V} = \frac{Q_{wo}}{\frac{\Delta V_{wo}}{\kappa}} = \frac{\kappa Q_{wo}}{\Delta V_o}$$

but this is just

$$C = \kappa C_{wo} \quad (35.7)$$

For a parallel plate capacitor, we have

$$C = \kappa \epsilon_o \frac{A}{d} \quad (35.8)$$

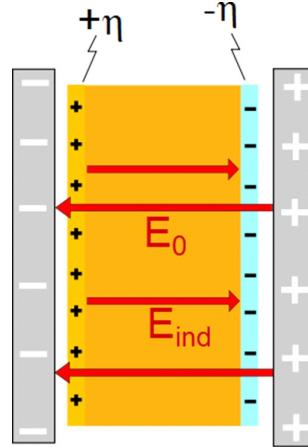
So where do you find values for κ ? For this class, we will look them up in the tables in books or on the internet. Here are a few values for our use.

Material	κ	Material	κ
Vacuum	1.00000	Paper	3.7
Dry Air	1.0006	Waxed Paper	3.5
Fused quartz	3.78	Polystyrene	2.56
Pyrex glass	4.7 – 5.6	PVC	3.4
Mylar	3.15	Teflon	2.1
Nylon	3.4	Water	80

Induced Charge

In the last discussion we discovered that if we put a dielectric inside a capacitor, we end up with polarized charges with the net result that there will be excess negative charge near the positive plate of the capacitor, and excess negative charge near the positive plates of the capacitor. In the middle of the dielectric, the charges are polarized in each atom, but still for any volume inside, the net charge is zero. The excess charge near each plate we will call the *induced charge*.

Since we have an induced positive charge on one side and an induced negative charge on the other side, we expect there will be an electric field directed from the positive to negative charge inside the dielectric.



The total field inside the dielectric is

$$E = E_{wo} - E_{ind} \quad (35.9)$$

where E_{wo} is the field due to the capacitor plates without the dielectric. From our previous discussion, we recall that

$$\Delta V = \frac{\Delta V_{wo}}{\kappa}$$

and we recall that the magnitude of the potential difference is given by

$$\Delta V = Ed$$

Then our new net field can be found

$$Ed = \frac{E_{wo}d}{\kappa}$$

or

$$E = \frac{E_{wo}}{\kappa}$$

and, recalling for a parallel plate capacitor (near the center) the field is approximately

$$E = \frac{\eta}{\epsilon_o}$$

then

$$E = E_o - E_{ind}$$

gives

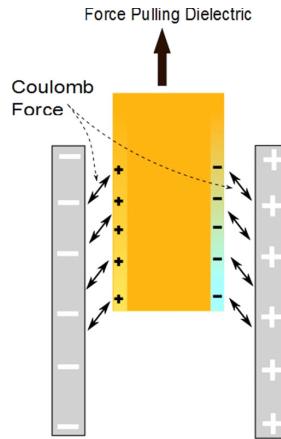
$$\frac{\eta}{\kappa\epsilon_o} = \frac{\eta}{\epsilon_o} - \frac{\eta_{ind}}{\epsilon_o}$$

and we can find the induced surface charge density as

$$\eta_{ind} = \eta \left(1 - \frac{1}{\kappa} \right)$$

You might guess that the induced charge is attracted to the charge on the plates, so a force is required (and work is required) to remove the dielectric once it is in place. If

we draw out the dielectric, we can see that the weaker field outside the capacitor causes little induced charge, but the stronger field inside the capacitor causes a large induced charge. A net inward force will result.



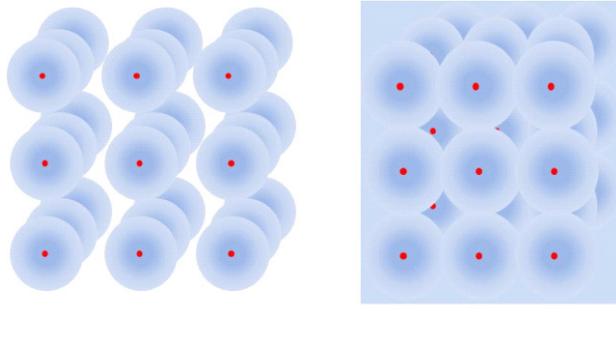
Electric current

Question 223.35.4

Question 223.35.5

For some time now, we have been talking about charge moving. We have had charge move from a battery to the plates of a conductor. We have had charge flow from one side to another of a conductor, etc. It is time to become more exact in describing the flow of charge. We should take some time to figure out why charge will move.

Let's consider a conductor again.



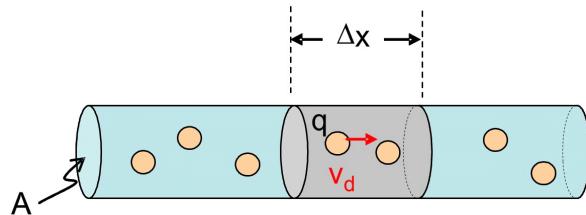
Insulator

Conductor

We remember that in the conductor, the valence electrons are free to move. In fact, they do move all the time. The electrons will have some thermal energy just because the conductor is not at absolute zero temperature. This thermal energy causes them to move

in random directions. (think of air molecules in a room).

Let's take a piece of a wire Δx long. The speed of the electrons along the wire (in the x -direction in this case) is called the *drift speed*, v_d , because the electrons just drift from place to place with a fairly small speed. This drift speed could be due mostly to thermal energy, so it can be very small or even zero (if no electric potential is applied). Of course, v_d , must be an average, each charge carrier will be moving random directions with slightly different speeds, so the x -component of the velocity will be different for each charge carrier, but on average they will move at a speed v_d .



So we will suppose that there are charge carriers of charge q_c that are moving through the wire with velocity v_d . Then we can write some length of wire, Δx , as

$$\Delta x = v_d \Delta t$$

The volume of the shaded piece of wire is

$$V = A \Delta x$$

Question 223.35.6

if there are

$$n = \frac{\#}{V}$$

charge carriers per unit volume, a *volume charge carrier number density*, then the total charge in our volume is

$$\Delta Q = n A \Delta x q_c$$

If we have electrons as our charge carrier, then q_c is just q_e .

We can substitute for Δx

$$\Delta Q = n A v_d \Delta t q_e$$

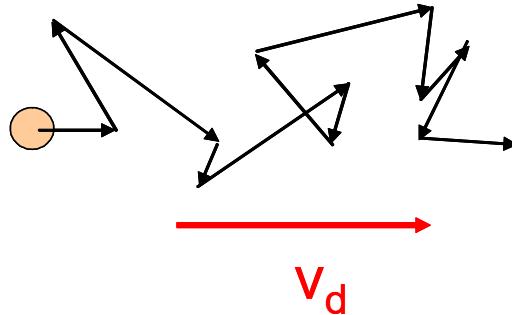
This gives the charge within our small volume. But it would be nice to know how much charge is going by, because we want moving charge. We can divide by Δt

$$\frac{\Delta Q}{\Delta t} = n A v_d q_e \quad (35.10)$$

to get a charge flow rate. This is very like our volume or mass flow rate in fluid flow.

We have an amount of charge going by in a time Δt .

I gave the flow velocity a special name, v_d . But I did not give all the reasons for using an average x -component of the velocity. If we think about it, we will realize that the electrons don't really flow in a straight line. They continually bump into atoms³⁰. So the actual path the electrons take looks more like this.

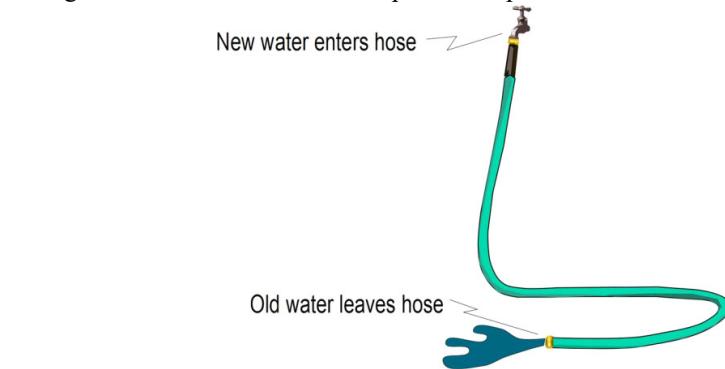


We only care about the forward part of this motion. It is that forward component that we call the *drift speed* of the electrons. It is much slower than the actual speed the electrons travel, and it depends on the type of conductor we are using.

We already know the name for the flow rate of charge, it is the electric current.

$$\frac{\Delta Q}{\Delta t} = I \quad (35.11)$$

We should take a minute to think about what to expect when we allow charge to flow. Think of a garden hose. If the hose is full of water, then when we open the faucet, water immediately comes out. The water that leaves the faucet is far from the open end of the hose, though. We have to wait for it to travel the entire length of the hose. But we get water out of the hose immediately! Why? Well, from Pascal's principle we know that a change in pressure will be transmitted uniformly throughout the fluid. This is like your hydraulic breaks. The new water coming in causes a pressure change that is transmitted through the hose. The water at the open end is pushed out.

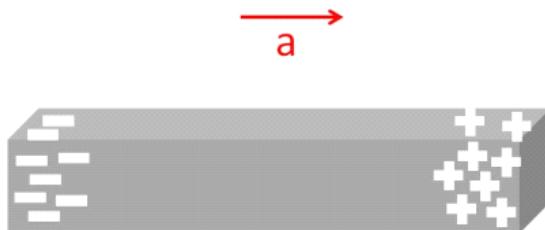


³⁰ We will refine this picture in the next lecture.

Question 223.35.7

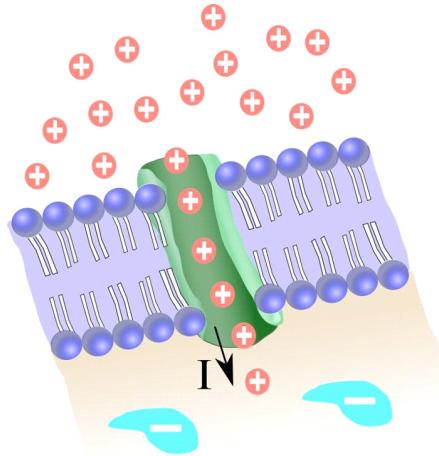
Current is a little bit like this. When we flip a light switch, the electrons near the near the switch start to flow at v_d . But there are already free electrons in all the wire. These experience a Pascal's-principle-like-push that makes the light turn on almost instantly.

There is a historical oddity with current flow. It is that the current direction is the direction positive charges would flow. This may seem strange, since in good conductors, we have said that electrons are doing the flowing! The truth is that it is very hard to tell the difference between positive charge flow and negative charge flow. In fact, only one experiment that I know of shows that the charge carriers in metals are electrons.



That experiment accelerates a conductor. The experiment is easier to perform using a centrifuge, but it is easier to visualize with linear motion. If we accelerate a bar of metal as shown in the preceding figure, the electrons are free to move about in the metal but the nuclei are all bound together. If the nuclei are accelerated they must go as a group. But the electrons will tend to stay with their initial motion (Newton's first law) until the end of the bar reaches them. At this point they must move because the electrical force of the mass of nuclei will keep them bound to the whole mass of metal. But the electrons will pile up at the tail end of the bar—that is—if it is the electrons that are free. When this experiment is performed, it is indeed the electrons that pile up at the tail end, and the forward end is left positive. This can be measured with a voltmeter.

Ben Franklin chose the direction we now use. He had a 50% chance if getting the charge carrier right. All this shows just how hard it is to deal with all these things we can't see or touch. And even more importantly, in semiconductors and in biological systems, it *is* positive charge that flows. In many electrochemical reactions *both* positive and negative charges flow. We will stick with the convention that the current direction is the direction that positive charges would flow regardless of the actual charge carrier sign.



Flow of positive charge through a gate into a neural cell.

Basic Equations

Voltage if a dielectric is placed between the plates of the capacitor (equation 35.6)

$$\Delta V = \frac{\Delta V_o}{\kappa}$$

Capacitance increases (equation 35.6)

$$C = \kappa C_o$$

For parallel plate capacitors we get

$$C = \kappa \epsilon_o \frac{A}{d}$$

The induced field in a dielectric is (equation 35.9)

$$E = E_o - E_{ind}$$

Current is the rate of charge flow (equation 35.11)

$$\frac{\Delta Q}{\Delta t} = I$$

Definition of current (equation 35.10)

$$I = nAv_dq_c$$

36 Current, Resistance, and Electric Fields

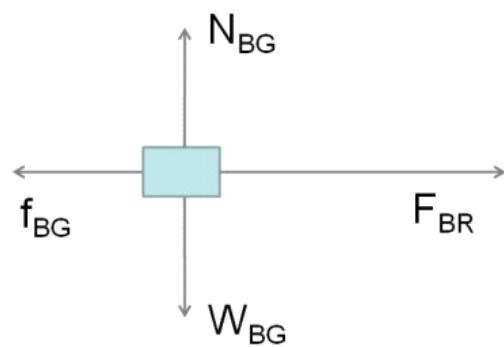
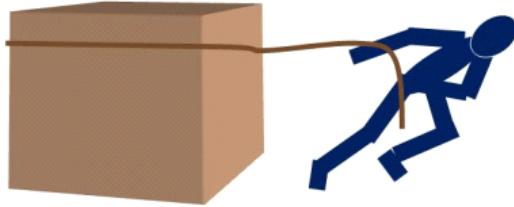
Fundamental Concepts

- There is a nonconservative (friction-like) force involved in current flow called *resistance*.
- A nonuniform charge distribution creates an electric field, which provides the force that makes current flow
- Current flow direction is defined to be the direction positive charge carriers would go
- The current density is defined as $J = nq_e v_d$
- Charge is conserved, so in a circuit, current is conserved.

Current and resistance

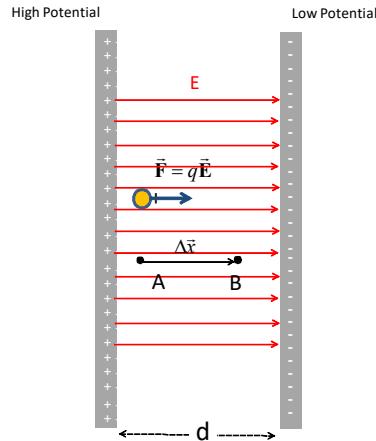
Question 223.36.1

We now have flowing charges, but our PH121 or Dynamics experience tells us that there is more. If we push or pull an object, we expect that most of the time there will be dissipative forces. There will be friction.



Question 223.36.2

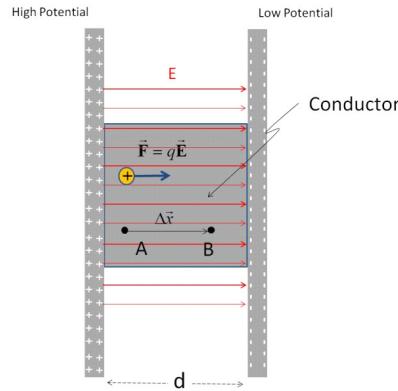
We should ask, is there a friction involved in charge movement? We already know how to push a charge, we use an electric field



The force is

$$F = qE$$

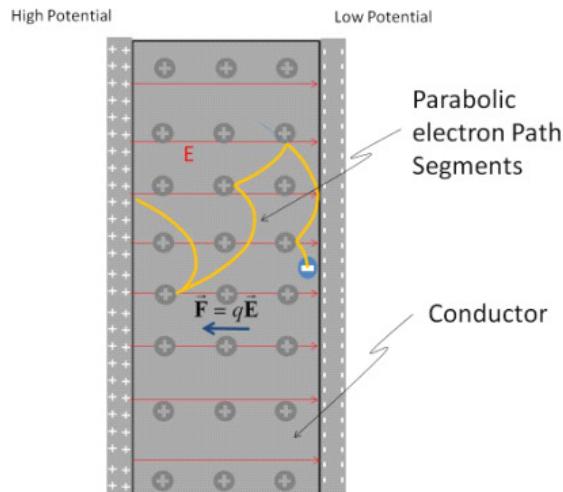
If we push or pull a box, it will eventually come to rest. In our capacitor there are no resistive forces for our charge to encounter. But suppose we place a conductor inside our capacitor, hooked to both plates



Question 223.36.3

Of course, in conductors we now know the charge carrier is an electron and it is negative, so let's try to redraw this picture to show the actual charge motion.

Now the charge is free to move inside of the conductor, but it is not totally unencumbered. The free charges will run into the nuclei of the atoms. The charges will bounce off. So as they travel through the material we will expect to see some randomness to their motion. This is compounded by the fact that the electrons already have random thermal motion. So the path the charge takes looks somewhat like this



We can recognize that each path segment after a collision must be parabolic because the acceleration will be constant

$$F = ma = qE$$

so

$$a = \frac{qE}{m}$$

we can describe the electron motion using the two of the kinematic equations

$$\begin{aligned} x_f &= x_i + v_{ix}\Delta t + \frac{1}{2}a_x\Delta t^2 \\ v_{fx} &= v_{ix} + a_x\Delta t \end{aligned}$$

and the path will be

$$x_f = x_i + v_{ox}\Delta t + \frac{1}{2} \left(\frac{qE}{m} \right) \Delta t^2$$

which is parabolic.

Of course, this is just for one electron, and only for a segment between collisions. We will have millions of electrons, and therefore, many millions of bounces. But for each electron, between bounces we expect a parabolic path. For considering current flow, we don't care about motion perpendicular to the current direction. So we can look only at the component of the motion in the flow direction. The net flow in the current direction is toward the positive plate. Let's see how this works.

Question 223.36.4

If we average the velocities of all the electrons we find

$$\begin{aligned} v_d &= \bar{v}_x \\ &= \bar{v}_{ix} + a_x\bar{\Delta t} \end{aligned}$$

the first term $\bar{v}_{ix} = 0$ because the initial velocities are random from the thermal and scattering processes. That is, on average, the electrons have no preferred direction after a bounce. This leaves

$$v_d = \left(\frac{qE}{m} \right) \bar{\Delta t}$$

The average time between collisions, $\bar{\Delta t}$, is sometimes given the symbol τ . Let's use this. Recall that current is

$$I = nAv_dq_c$$

Then

$$v_d = \left(\frac{q\tau}{m} \right) E$$

and we can write our current equation as

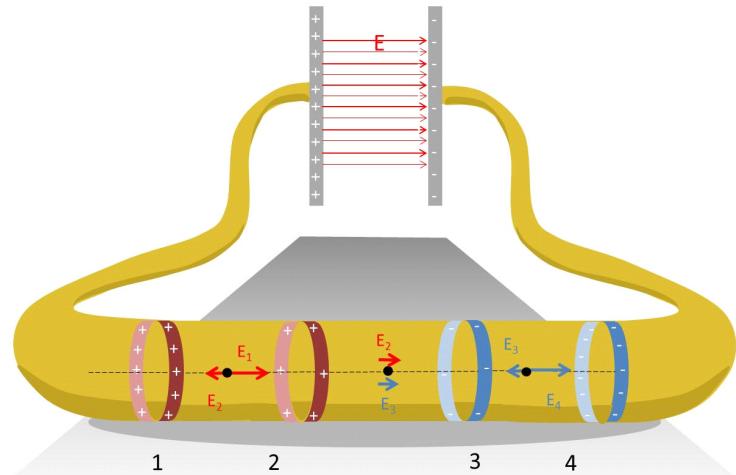
$$I = nqA \left(\frac{q\tau}{m} \right) E$$

We have shown that the current is directly proportional to the field inside the conductor.

Question 223.36.5

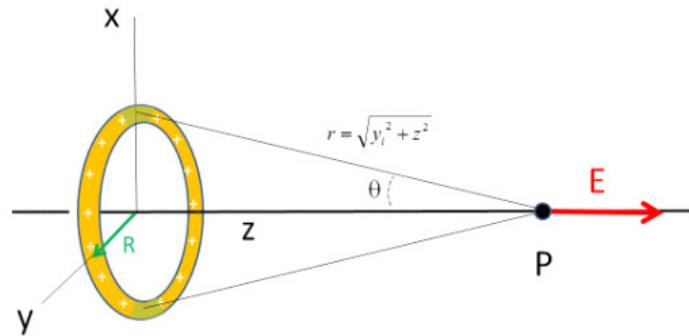
It is this field that causes the charges to flow.

But let's look even closer. Suppose we connect our two plates with a wire instead of filling their gap with a conductor. If current flows through the wire, there must be a field in the wire. But how does it get started?



This figure is supposed to show our wire connected to the capacitor. The capacitor is in the background, and the wire loops close to us. The end of the wire that is connected to the positive side of the capacitor will become positively charged, and the end connected to the negative side of the capacitor will become negatively charged. But if we look at the wire an infinitesimal time after the connection has happened, the wire will not be uniformly charged. It will take some time for the charges to reach equilibrium. In the mean time, the charge is stronger near the plates, and diminishes toward the middle.

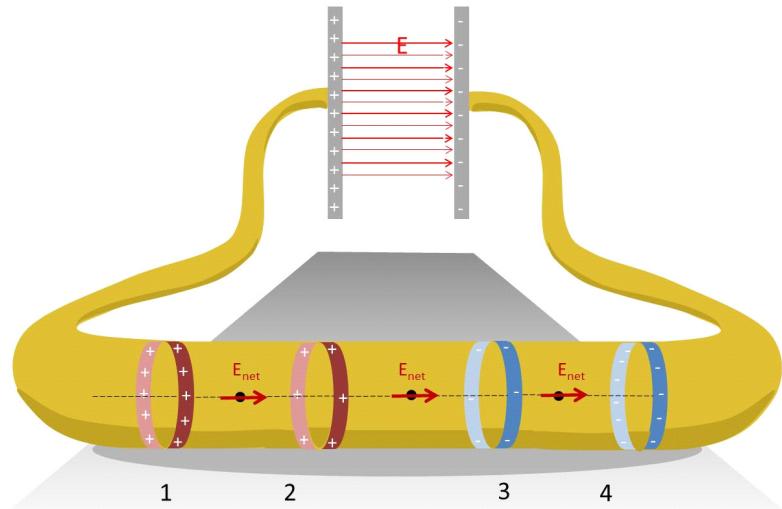
We can't find the exact field in the conductor without resorting to a computational solution, but we can mentally model the situation by viewing the wire as consisting of rings of charge that vary in linear charge density. We know the field along the axis due to a ring of charge because we have done this problem in the past.



$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{zQ}{(R^2 + z^2)^{\frac{3}{2}}} \hat{k}$$

We know the field is along the axis and that it diminishes with distance from the ring.

Now consider the field due to ring 1. As we move to the right, away from ring 1 that field will diminish with distance. Also consider the field due to ring 2. As we move to the right toward ring 2 the field due to ring two will grow. The field due to ring two grows at the same rate that the field from ring 1 diminishes. The fields 1, and 2 add up to a constant value along the axis for every point in between the two rings. Now consider the field on the right side of ring 2 and the field on the left side of ring 3. A little thought shows that the situation is the same as that for rings 1 and 2. We will have a constant net field between the two rings.



Likewise for the region between rings 3 and 4. There is a constant net electric field at all points along the wire. This field points from positive to negative. It will exert a force

$$F = qE_{net}$$

on the free charges *inside* the wire. These free charges are not extra charge. They are the free electrons that are loosely attached to the metal atoms that make up the wire. So these free charges are distributed throughout the volume of the wire. These free charges will accelerate, forming a current inside the wire.

Note that these free charges are not just on the surface, they are inside the wire, even on the axis of the wire in the center. We no longer have a static equilibrium, so we no longer have excess charge only on the surface.

All this usually happens very fast, so when we switch on a light, we don't notice the time it takes for the current to start. But this uneven distribution of charge is the reason we get a current.

Current density

Question 223.36.6

We now realize that when there is an electric field inside a wire, there will be current flow inside the wire. The flow goes through the volume of the wire.

The rate of flow is given by

$$I = \frac{\Delta Q}{\Delta t} = nq_e A \left(\frac{q_e \tau}{m_e} \right) E$$

for steady current flow. Here we are writing $q = q_e$ for the electron charge and $m = m_e$ for the electron mass, since our charge carrier is an electron..

The unit for current flow is

$$\frac{\text{C}}{\text{s}} = \text{A}$$

Question 223.36.7

where A is the symbol for an *Ampere* or, for short, an *amp*.

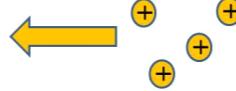
Historically there was no way to tell whether negative charges were flowing or whether positive charges were flowing. It really did not matter so much in the early days, since a flow of positive charges one way is equivalent to a flow of negative charges the other way.

Case 1: Negative charges flow to the right



Result: Left side is more positive than before,
Right side is more negative than before

Case 2: Negative charges flow to the right

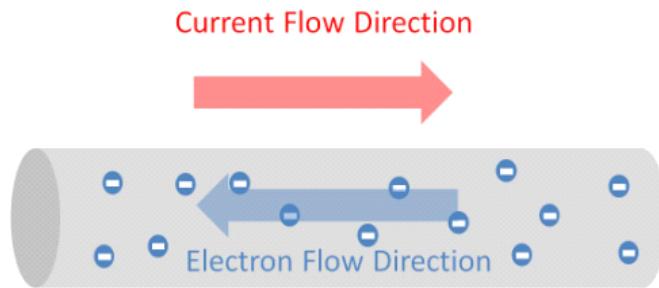


Result: Left side is more positive than before,
Right side is more negative than before

Worse, we know that for some systems there are positive charge carriers and for others negative charge carriers.

By convention, we assign the direction of current flow as though the charge carrier were positive.

This is great for biologists, where the charge carriers are positive ions. But for electronics this gives us the uncomfortable situation that the actual charge carriers, electrons, move in the direction opposite to that of the current.



Let's look again at our definition of current

$$I = \frac{\Delta Q}{\Delta t} = nq_e A \left(\frac{q_e \tau}{m_e} \right) E$$

If we, once again, write this in terms of v_d

$$v_d = \left(\frac{q \tau}{m} \right) E$$

then after rearranging, we have

$$I = (nq_e v_d) A$$

Question 223.36.8

the part in parentheses contains only bulk properties of the conductor material, the number of free charges, the charge of the charge carrier, and the drift speed which depends on the material structure of the conductor. The final factor is just the cross sectional area of the wire. It gives the geometry of the wire we have made out of the bulk material (say, copper). It is convenient to group all the factors that are due to bulk material properties

$$J = nq_e v_d$$

then the current would be

$$I = JA$$

Note how similar this is to a surface charge density

$$Q = \eta A$$

For a static charged surface, Q is the surface charge density multiplied by the particular area. For our case we have a total current, I that is the material properties multiplied by an area. By analogy we could call this new quantity, J , a kind of density, but now our charges are moving. So let's call it the *current density*.

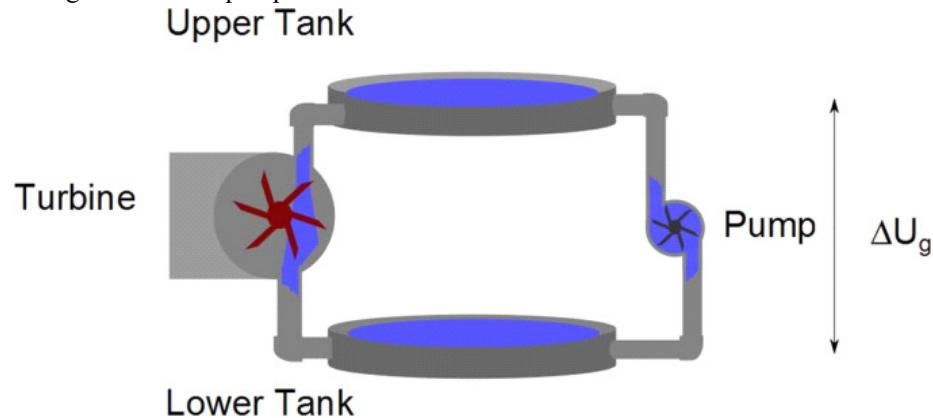
Notice that it is the cross sectional area of the wire that shows up in our current

equation. This is another indication that the charge is not flowing along the surface, but that it is deep within the wire as it flows.

Conservation of current

Question 223.36.9

Let's go back to our pumps and turbines.

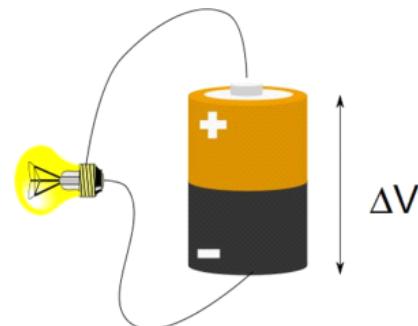


How much of the water is “used up” in turning the turbine? Another way to say this is to ask if there are 20l of water entering the turbine, how much water leaves the turbine through the lower pipe?

Question 223.36.10

If the turbine leaks, then we might lose some water, but if all is going well, then you can guess that 20l of water must also leave the turbine. We can't lose or gain water as the turbine is turned. But we must be losing something! We must be giving up something to get useful work out of the system. That something that we lose is potential energy.

Now consider a battery. How much of the current is “used up” in making the light bulb light up?



This case is really the same as the water case. The electric current is a flow of electrons. The flow loses potential energy, but we don't create or destroy electrons as we convert the potential energy of the battery to useful work (like making light) just like we did not create or destroy water in making the turbine turn.

But surely the water slowed down as it traveled through the turbine—didn't it? Well, no, if the water slows down as it goes through the turbine, then the pipe below the turbine would run dry. This does not happen. The flow rate through a pipe does not change under normal conditions, and under abnormal conditions, we would destroy the pump or the turbine! If we throw rocks off a hill, they actually gain speed when the water loses potential energy. Now the flow rate is slower with a turbine in the pipe than it would be with no turbine in the pipe! But with the turbine in the pipe, the flow rate is the same throughout the whole pipe system.

Like the water case, the flow rate of charge does not change from point to point in the wire. The same amount of charge per unit time leaves the wire as went in.

This explains the reasoning behind one of the great laws of electronics

The current is the same at all points in a current-carrying wire.

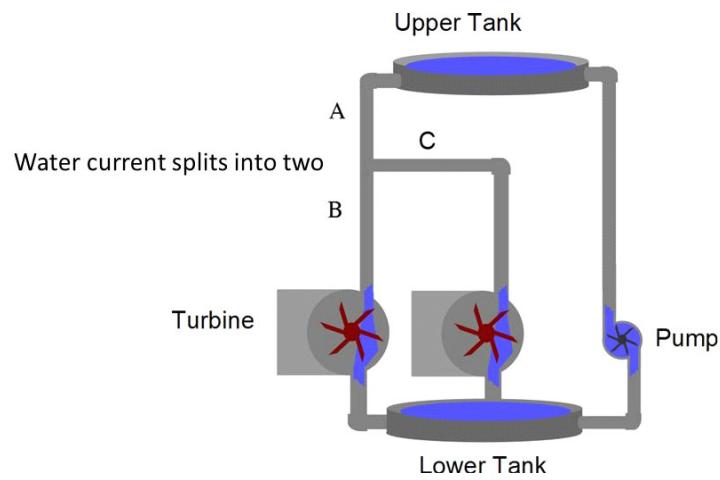
Like in the water case, the electrons would flow faster if there were no light bulb and just a continuous wire. We can have different flow rates in our wire depending on how much resistance there is to the flow. But the flow rate will be the same in all parts of the wire system.

This leads to the second of the pair of rules called Kirchhoff's laws:

$$\sum I_{in} = \sum I_{out}$$

If the wire branches into two or more pieces, the current will divide. This is not too surprising. The same is true for water in a pipe

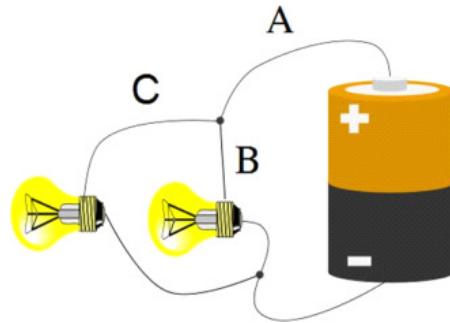
Question 223.36.11



In the figure the flow through pipe segment A is split into two smaller currents that flow through pipe segments B and C . We would expect that the flow through B and C combined

Question 223.36.12 must be equal to the flow through A .

The same must be true for electrical current. The situation is shown in the next figure.



The current that flows through wires B and C combined must be equal to the current that came through wire A .

Question 223.36.13 Basic Equations

37 Ohm's law

Fundamental Concepts

- The material property of a conductor that tells us how well the conductor material will allow current to flow through it is called the conductivity
- The inverse of conductivity is the resistivity
- Resistivity may be temperature dependent
- Resistance depends on the resistivity of the material and the geometry of the conductor piece. For a wire it is given by $R = \rho A/L$
- For many conductors, the change in voltage across the conductor is proportional to the current and the resistance. This is called Ohm's law
- The ideal voltage delivered by a battery is called the "emf" and is given the symbol \mathcal{E}
- Some materials do not follow Ohm's law. They are called nonohmic
- The Earth has a magnetic field
- Magnets have "magnetic charge centers" called poles and there is a magnetic field.
- Magnetic poles don't seem to exist independently

Conductivity and resistivity

Question 223.37.1

We defined the current density last lecture

$$J = nq_e v_d$$

but we know that the drift speed is

$$v_d = \left(\frac{q_e \tau}{m_e} \right) E$$

so we can write the current density as

$$\begin{aligned} J &= nq_e \left(\frac{q_e \tau}{m_e} \right) E \\ &= \left(\frac{nq_e^2 \tau}{m_e} \right) E \end{aligned}$$

The factor in parentheses depends only on the properties of the conducting material.

For example, if the material is copper, then we would have the $n_{copper} = 8.5 \times 10^{28} \frac{1}{\text{m}^3}$ as the number of valence electrons per unit meter cubed for copper. The mean time between collisions is something like $\tau_{copper} = 2.5 \times 10^{-14} \text{ s}$. So our quantity in parentheses is

$$\begin{aligned}\left(\frac{nq_e^2\tau}{m_e}\right) &= \frac{\left(8.5 \times 10^{28} \frac{1}{\text{m}^3}\right) \left(1.6 \times 10^{-19} \text{ C}\right)^2 \left(2.5 \times 10^{-14} \text{ s}\right)}{9.11 \times 10^{-31} \text{ kg}} \\ &= 5.9715 \times 10^7 \frac{\text{A}^2}{\text{m}^3} \frac{\text{s}^3}{\text{kg}} \\ &= 5.9715 \times 10^7 \frac{1}{\Omega \text{ m}}\end{aligned}$$

The field is due to something outside of the conducting material (e.g. the battery). Notice that again we have grouped all the properties of the material together. Lets give a name to the quantity in parentheses that contains all the material properties. Since this quantity tells us how easily the charges will go through the conductive material, we can call this the *conductivity* of the material.

$$\sigma = \frac{nq_e^2\tau}{m_e}$$

Then

$$J = \sigma E$$

The current density depends on two things, how well the material can allow the current to flow (bulk material properties related to conduction), σ , and the field that motivates the current to flow, E .

The current, then, depends on these two items, as well as the cross sectional area of the wire

$$\begin{aligned}I &= JA \\ &= \sigma EA\end{aligned}$$

Really, the conductivity is more complicated than it appears. The mean time between collisions, τ , depends on the structure of the conductor. Different crystalline structures for the same element will give different values. Think of trying to walk quickly through the Manwering Center crowds during a class break. This takes some maneuvering. But if all the people were placed at equally spaced, regular intervals, it might be easier to make it through quickly. It would also be easier if the crowd stood still. Likewise, the position of the atoms in the conductor make a big difference in the conductivity, and thermal motion of those atoms also makes a large difference. We would expect the conductivity to depend on the temperature of the material.

Resistivity

Question 223.37.2

It is common to speak of the opposite of the concept of conductance. In other words, how hard it is to get the electrons to travel through the conductive material. For example, we might want to build a heating device, like a toaster or space heater. In this case, we want friction in the wires, because that friction will produce thermal energy. So specifying a conductive material by how much friction it has is useful. How much the material impedes the flow of current is the opposite of how much the material allows the current flow, so we expect this new quantity to be the inverse of our conductivity

$$\rho = \frac{1}{\sigma} = \frac{m_e}{nq_e^2\tau}$$

Special conductors are often made that use “impurities,” that is, trace amounts of other atoms, to increase or decrease the resistivity of those conductive materials. The thermal dependence can be modeled using the equation

$$\rho = \rho_o (1 + \alpha (T - T_o))$$

where ρ_o is the resistivity at some reference temperature (usually 20 °C) and α is a constant that tells us how our particular material changes resistance with temperature. It is kind of like the specific heat in thermodynamics $Q = C\Delta T$. This is an approximation. It is a curve fit that works over normal temperatures. But we would not expect the same resistive properties, say, if we melt the material. The position of the atoms would change if the material goes from solid to liquid. So we will need to be careful in how we use this formula.

Here are some values of the conductivity, resistivity, and temperature coefficients for a few common conductive materials.

Question 223.37.3

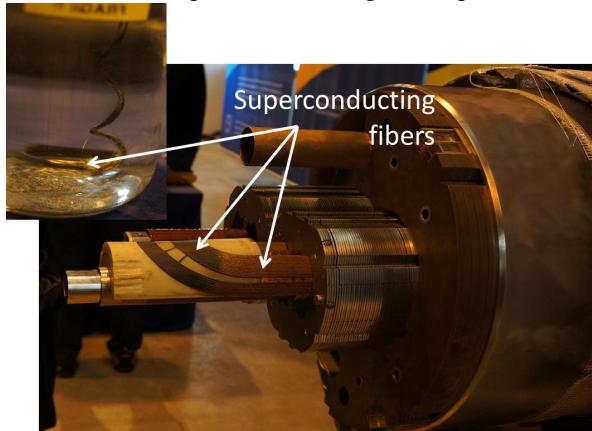
Material	Conductivity ($\Omega^{-1} m^{-1}$)	Resistivity (Ωm)	Temp. Coeff. (K^{-1})
Aluminum	3.5×10^7	2.8×10^{-8}	3.9×10^{-3}
Copper	6.0×10^7	1.7×10^{-8}	3.9×10^{-3}
Gold	4.1×10^7	2.4×10^{-8}	3.4×10^{-3}
Iron	1.0×10^7	9.7×10^{-8}	5.0×10^{-3}
Silver	6.2×10^7	1.6×10^{-8}	3.8×10^{-3}
Tungsten	1.8×10^7	5.6×10^{-8}	4.5×10^{-3}
Nichrome	6.7×10^5	1.5×10^{-6}	0.4×10^{-3}
Carbon	2.9×10^4	3.5×10^{-5}	-0.5×10^{-3}

Superconductivity

The relationship

$$\rho = \rho_o (1 + \alpha (T - T_o))$$

also breaks down at low temperatures. The low end is very important these days. For some special materials, the resistivity goes to zero when the material is cold enough. We call these materials superconductors. A superconductor can carry huge currents, because there is no loss of energy, and no heat generated without any friction. Unfortunately most superconducting materials only operate at temperatures near absolute zero. But a few "high temperature" superconductors operate at temperatures as high as 125 K. This is still very cold (-150°C), but these temperatures are achievable, so some superconducting products are possible. As you can guess, there is very active research in making superconductors that operate at even higher temperatures.



Superconducting fiber material and superconducting magnet at CERN. These superconductors operate at 1.9K.

Ohm's law

Let's pause to review, Current density is given by

$$J = \sigma E$$

or now by

$$J = \frac{1}{\rho} E$$

Then the current is given by

$$\begin{aligned} I &= JA \\ &= \frac{A}{\rho} E \end{aligned}$$

If the field is similar to our capacitor field, nearly uniform in our conducting wire, then the potential would be just

$$\begin{aligned} \Delta V &= Ed \\ &= E\Delta s \end{aligned}$$

and then the electric field is approximately given by

$$E = \frac{\Delta V}{\Delta s}$$

For our wire of length L this is

$$E = \frac{\Delta V}{L}$$

Then we can use this field to write our current

$$I = \frac{A}{\rho} \frac{\Delta V}{L}$$

Once again, let's group together all the structural and material properties of the wire.

We have

$$I = \left(\frac{A}{L\rho} \right) \Delta V$$

or with a little algebra,

$$\Delta V = I \left(\rho \frac{L}{A} \right)$$

The part in parenthesis contains all the friction terms. It says that the longer the wire, the more friction we will experience. This makes sense. If you are familiar with fluid flow. The longer the hose, the more resistance. It also says that the larger the area, the lower the friction. That is also reasonable, since the electrons will have more places to go unrestricted if the area is bigger. In water flow, the larger the pipe, the less the water interacts with the sides of the pipe and therefore the lower the friction. This situation is analogous.

Question 223.37.4

We should give a name to this quantity that describes the frictional properties of the wire. We will call it the *resistance* of the wire.

$$R = \rho \frac{L}{A}$$

so that we can write

$$I = \frac{\Delta V}{R}$$

The resistance has units of

$$\frac{V}{A} = \Omega$$

where Ω is given the name of *ohm* after the scientist that did pioneering work on

resistance.

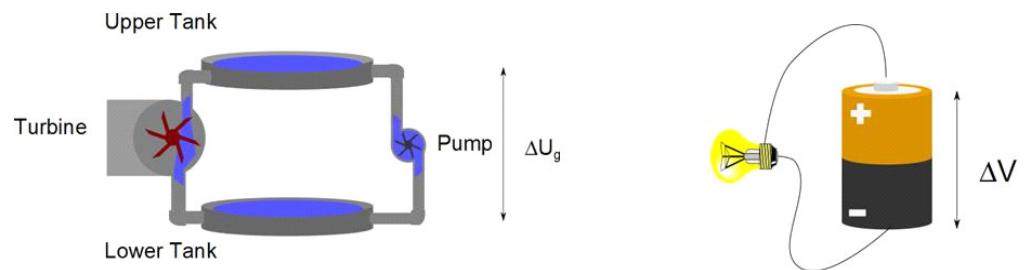
The relationship

$$I = \frac{\Delta V}{R}$$

is called *Ohm's law*.

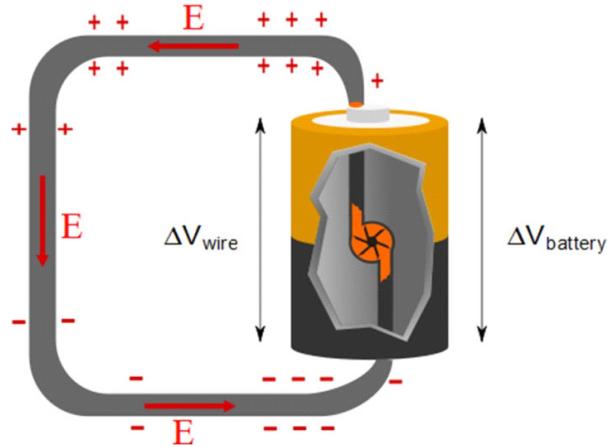
Life History of an electric current

Let's go back and think about our pump model for a battery.



The pump is a source of potential energy *difference*. This is what a battery does as well. The battery is a charge pump. It moves the charges from a low to a high potential. So it is a source of electric potential. The battery's job is to provide the charge separation that creates the electric field that drives the free charges, making the current.

A positive charge in the wire on the negative side of the battery is pumped up to the positive side through a chemical process. We can mentally envision a small charge pump inside of the battery



The battery is the source of the potential. A positive charge near the negative side of the battery would be pumped up to the positive side of the battery. It would gain potential energy

$$\Delta U_{battery} = q\Delta V_{battery}$$

Then it would “fall” down the wire. It must lose all of the potential energy it gained. So it will loose

$$|\Delta U_{wire}| = |\Delta U_{battery}|$$

But if the battery potential energy change is positive, the wire change must be negative. We can see that

$$\Delta V_{wire} = -\Delta V_{battery}$$

so the potential change in the wire is negative. We sometimes call this a potential “drop.”

The field forces our charge to move through this wire much like the gravitational field forces rocks to fall. The positive charge ends up at the negative end of the battery again, ready to be pumped up to make another round.

Of course, really this process goes backwards in electrical circuits, since our charge carriers are negative, but we recall that mathematically negative charges going the opposite way is the same. So we will make this picture our mental model of a current.

Emf

We have ignored something in our pump model of a battery. In real water flow, there

would be resistance to the flow even inside the pump. This resistance would be small, but not zero. So the actual potential energy gain would be

$$\Delta U = \Delta U_{\text{ideal}} - U_{\text{loss due to friction in the pump}}$$

The same is true for an actual battery. There is some resistance in the battery, itself.

$$\Delta V = \Delta V_{\text{ideal}} - \Delta V_{\text{loss due to resistance in the battery}}$$

Now that we have Ohm's law, we can see what $\Delta V_{\text{loss due to resistance in the battery}}$ would be in terms of the internal resistance of the battery and the current that flows. Referring to the last figure, there is only one way for the current to go. So for this circuit, the current must be the same throughout the entire circuit, even in the battery! If we call the small resistance in the battery r , then

$$\Delta V_{\text{loss due to resistance in the battery}} = Ir$$

then the actual potential energy provided by the battery is

$$\Delta V = \Delta V_{\text{ideal}} - Ir$$

It is traditional to give the ideal voltage a name and a symbol. And we have already encountered this name. It is “emf.” Recall that at one time, the letters ‘e’, ‘m’, and ‘f’ stood for something. But not any more. It is just a name. It is pronounced “ē-em-ef,” and the symbol is a script capital \mathcal{E} . So we can write

$$\Delta V = \mathcal{E} - Ir$$

Sometimes you will hear \mathcal{E} referred to as the voltage you would get if the battery is not connected (the “open circuit” voltage). This is the voltage marked on the battery. Notice that the actual voltage provided at the battery terminals depends on how much current is being drawn from the battery. So if you are draining your battery quickly (say, using your electric starter motor to start your car engine) the voltage supplied by your battery might drop (your lights might dim while the starter motor runs). You are not getting 12 V because the current I is large while the starter motor runs. We will change to this new symbol for ideal voltage. But we should keep in mind that actual voltages delivered may be significantly less than this ideal emf unless we plan our designs carefully.

Ohmic or nonohmic

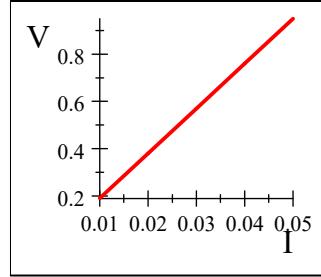
Question 223.37.5

This simple model of resistance is great for understanding simple things. Wires, and resistors do work like this. If we were to take a set of measurements of ΔV and I , we

expect a straight line

$$\begin{aligned}y &= mx + b \\ \mathcal{E} &= \Delta V = RI + 0\end{aligned}$$

where R is the slope.



But there are times when the model fails terribly. An incandescent light bulb is an example that we can quickly understand. The resistance at any one moment fulfils Ohm's law

$$I = \frac{\mathcal{E}}{R}$$

but light bulbs get hot. The resistance will change in time. So our relationship is now time dependent. Starting with the resistivity,

$$\rho = \rho_o (1 + \alpha (T - T_o))$$

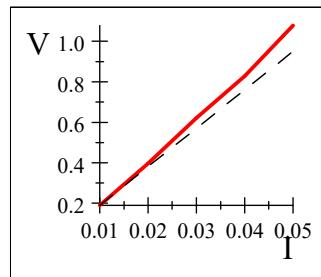
let's multiply both sides by A/L .

$$\frac{A}{L} \rho = \frac{A}{L} \rho_o (1 + \alpha (T - T_o))$$

this gives

$$R = R_o (1 + \alpha (T - T_o))$$

So if the resistance is temperature dependent, the slope of the line will change as we go along making measurements. We might get something like this



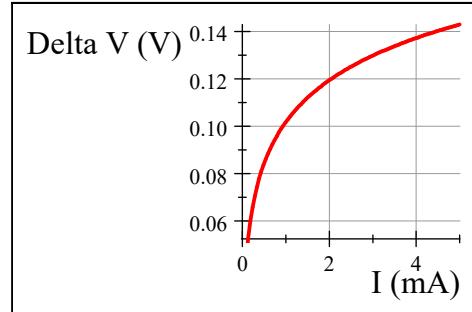
The dashed line is what we expect from Ohm's law. The solid line is what data from a light bulb would actually look like. We could use our temperature dependent resistance,

and realize that the temperature is a function of time, to obtain

$$I = \frac{\mathcal{E}}{R_o(1 + \alpha(T(t) - T_o))}$$

Since this set of measurements is not strictly following Ohm's law, we will say that the light bulb is *nonohmic*.

Many common circuit elements are vary nonohmic. A diode, for example, has a ΔV vs. I relationship that looks like this.



We can now understand how an electric current is formed. Hopefully you have taken, are taking, or will take ME210 so you will know how to build simple circuits with resistances and capacitances. But for this class, we will now investigate a new force, the magnetic force.

Power in resistors

We learned that the resistance in a resister depends on the temperature of the resister, and even have an approximate relationship that shows how this works

$$R = R_o(1 + \alpha(T - T_o))$$

so we know that temperature and resistance are related. But most of us have used a toaster, or an electric stove, or an electric space heater, etc. How does an electric circuit produce heat? or even light from a light bulb?

To answer this let's think of the energy expended as an electron travels a circuit. The potential energy expended is

$$\Delta U = q\Delta V$$

where the ΔV comes from the battery, so we could write this as

$$\Delta U = q\mathcal{E}$$

This is the energy lost as the electron travels from one side of the battery to the other.

We could describe how fast the energy is lost by dividing by the time it takes the electron to make the trip

$$\frac{\Delta U}{\Delta T} = \frac{q}{\Delta t} \mathcal{E}$$

but of course we want to do this for more than one electron. Let's take a small amount of charge, ΔQ , then

$$\begin{aligned}\frac{\Delta U}{\Delta T} &= \frac{\Delta Q}{\Delta t} \mathcal{E} \\ \frac{\Delta U}{\Delta T} &= \frac{\Delta Q}{\Delta t} \mathcal{E}\end{aligned}$$

and if we make the small group of charge very small we have

$$\frac{dU}{dT} = \frac{dQ}{dt} \mathcal{E}$$

and we recognize dU/dt as the power and dQ/dt as current, then

$$\mathcal{P} = I\mathcal{E}$$

This is the power supplied by the battery in moving the group of electrons through the circuit. But from conservation of energy, the charge packet must lose all the energy that the battery provides, so

$$\mathcal{P}_{battery} = \mathcal{P}_R = I\Delta V_R$$

is the energy that leaves the circuit as the packet of charge moves.

This works for any resistance

$$\mathcal{P}_R = I\Delta V_R$$

Then we can use Ohm's law

$$\Delta V_R = IR$$

to find

$$\begin{aligned}\mathcal{P}_R &= I(IR) \\ &= I^2 R\end{aligned}$$

But where does this energy go? This is the energy that makes the heat in the space heater, or the light in the light bulb.

Magnetism

Most people have used a magnet. at some time. They come as ads that stick to a refrigerator. They are the working part of a compass. They hold the pieces of travel games to their boards, etc. So I think we all know that magnets stick to metal things. But do they stick to all metal things?

The answer is no, only a few metals work. Iron and Nickel and Cobalt are some that do. Aluminum and Copper do not. By the time we are done studying magnetism, we should be able to explain this.

Magnets are very like charged objects in some ways. They can attract or repel each other. They attract “unmagnetized” materials. But there are some important differences.

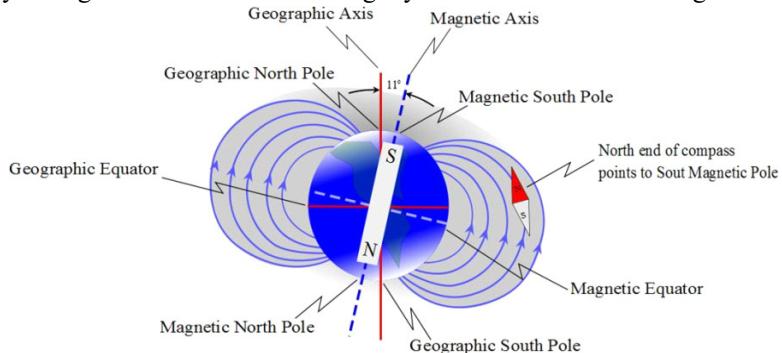
Bar Magnet Demo –
Make this like the
first charge demo

Bar Magnet Demo – Alternate, use the array of iron arrows and an overhead projector with the bar magnet

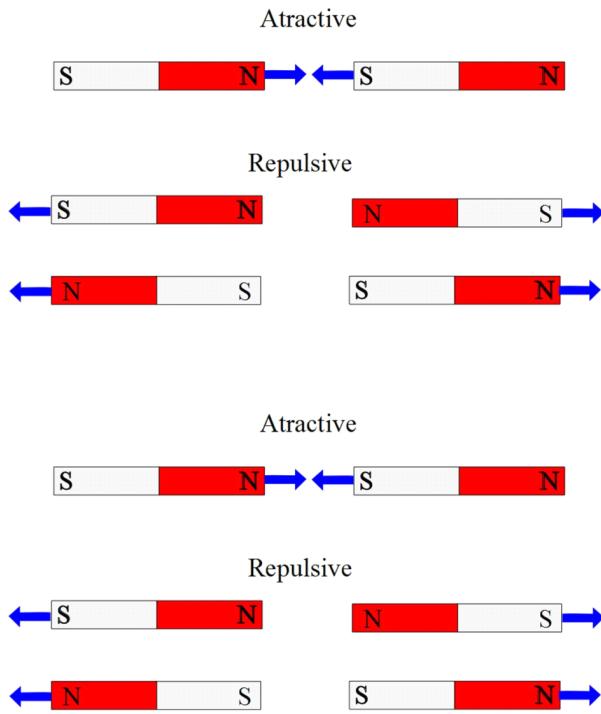
More Bar Magnet
Demo – Like Poles

Notice that a “magnetic charge” seems to be induced in some metal objects, but not in other common objects. This is very different than electric charge and electric polarization! And we should state explicitly that for magnets, there seem to be both “charges” in the same object! We call the “charge centers” the *poles* of the magnet. We find that one pole attracts one of the poles of a second magnet and repels the other. If we turn around the first magnet, we find that our pattern of attraction and repulsion reverses. Because magnets were used for centuries in navigational compasses, we call one pole the *north pole* of the compass and the other the *south pole* of the magnet. The north pole is the pole that would orient toward the north. Why does this happen?

I hope your high school science class taught you that the Earth has a magnetic field.



So we constantly live under the influence of a large magnet! Now let's hang both of our magnets from a string, and see which way they like to hang. The north facing end we will label *N* and the south facing end we will label *S*. Now we can see that the two *N* ends repel each other and the two *S* ends repel each other. But a *N* end and a *S* end will attract.

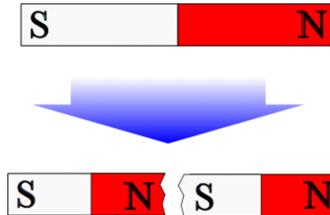


Once again we have a situation where we can define a mover object and an environmental object. We can picture one of the magnets making a magnetic field and the other magnet moving through this field. Of course both magnets make magnetic fields, but since a magnet can't make a magnetic field that moves itself, we won't draw this self-field for the mover magnet. We just draw the field for the environmental magnet. We did this in our Earth-compass picture. The Earth was the environmental magnet and the compass was the mover magnet.

One quirk of history is that since a *N* end of a magnet is attracted to the North part of the Earth. But north end of magnets are attracted to south poles of magnets, the Earth's geographic north pole must be a magnetic south pole!

One common misconception is that there is one specific place that is the magnetic north pole. Really it is a region near Newfoundland where the field strength actually varies quite a bit. You may have heard people discuss how the poles switch every so often. This is true, and we don't fully understand the mechanism for this.

There is a large difference between the magnetic force and the electric force. Electric charges are easy to separate. But magnetic poles are not at all easy to separate. If we break a magnet



we end up with each piece being a magnet complete with both north and south ends. This is very mysterious! something about the source of the magnetic field must be very different than for the source of the electric field. We will investigate the source of a magnetic field as we go.

The Earth's magnetic fields affects many biological systems. One of these is a bacteria that contain small permanent magnets inside of them to help them find the mud they live in.

In the 1990's there was a health fad involving magnets. Many people bought magnets to strap on their bodies. They were supposed to reduce aging and give energy. Mostly they stimulated the economy. But we will find that magnetic fields can alter the flow of blood (but these magnets did not do so, the FDA would not allow strong enough magnets to be sold as apparel to have this effect). Another common place to find magnetic fields is the MRI devices used in hospitals to make images of the interior of bodies.

Question 223.37.6

Basic Equations

38 Magnetic Field

Fundamental Concepts

Pass out magnets on sticks

Pass out magnets

We have now experience with two non-contact forces, the gravitational force and the electric or Coulomb force. In both cases, we have found that there is a field involved with the production of this force. We can guess that this is true for the magnetic force as well.

The discovery of this field involved an accidental experiment, and understanding this experiment gives us great insight into the nature of this field and where it comes from. So we will spend a little time describing it.

Fundamental Concepts in the Lecture

- A long wire that carries a current produces a magnetic field
- The magnetic field due to a long wither with current becomes weaker with distance and forms concentric cylinders of constant magnetic field strength
- The direction of the long-wire-with-current field is given by a right-hand-rule.
- The field due to a moving charge is given by the *Biot-Savart law*

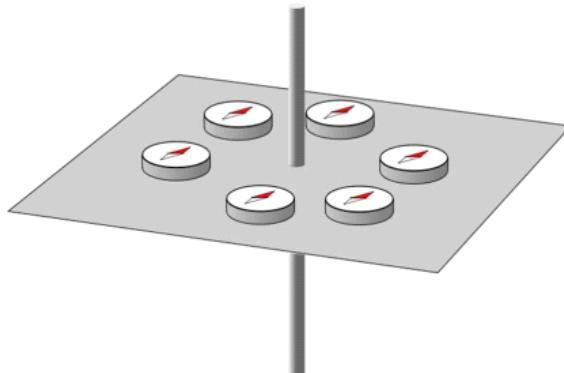
$$B = \frac{\mu_0}{4\pi} \frac{qv \sin \theta}{r^2}$$

Discovery of Magnetic Field

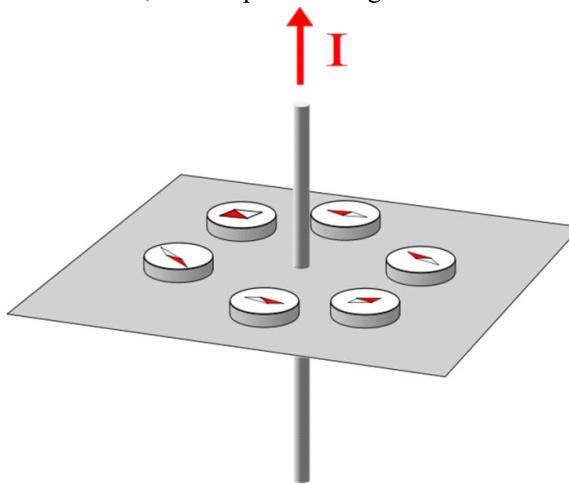
Question 223.38.1

In 1819 a Dutch scientist named Oersted was lecturing on electricity. He was actually making the point that there was no connection between electricity and magnetism. He had a large battery connected to a wire. A large current flowed through the wire. By chance, Oersted placed a compass near the wire. He had done this before, but this time the wire was in a different orientation than in previous demonstrations. To his great surprise, the compass needle changed direction when it was placed near the wire!

A similar experiment, but this time with several compasses, is shown in the next figure.



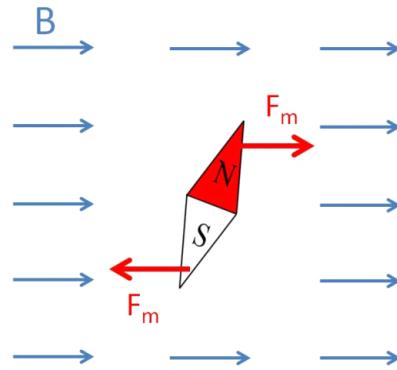
When the current is turned on, the compasses change direction.



This is a very good clue that there really *is* a connection between electricity and magnetism.

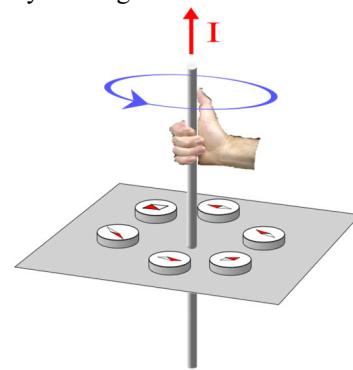
Oersted's Experiment Demo: Use the 106 boards and compasses

We know that a compass orients itself in the Earth's magnetic field. We can infer that the compass needle will orient in any magnetic field. In the next figure you can see that there is a force on each end of the needle due to the magnetic field.



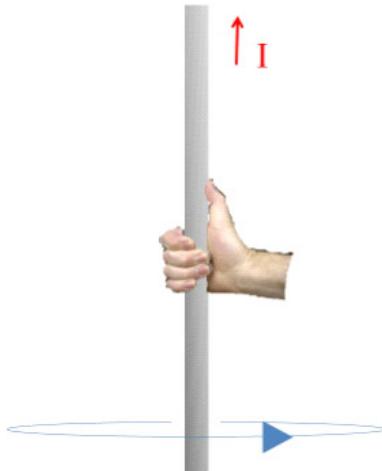
Notice that we have marked the environmental magnetic field with the letter B . This is traditional. Magnetic fields are often called B -fields for this reason. But more importantly, this looks very like an electric dipole in a constant electric field. We know enough about the dipole situation to predict that there will be a torque, and that there will be a stable equilibrium when the compass needle is aligned with the magnetic field.

Since our compasses oriented themselves near the current carrying wire, there must be a magnetic field caused by the current in the wire. The field shown in the last figure is uniform, but the field of our wire cannot be uniform. The compasses pointed different directions. A common way to describe this field is with a right-hand-rule. We imagine grabbing the wire with our right hand with our thumb pointing in the current direction. The field direction is given by our fingers.

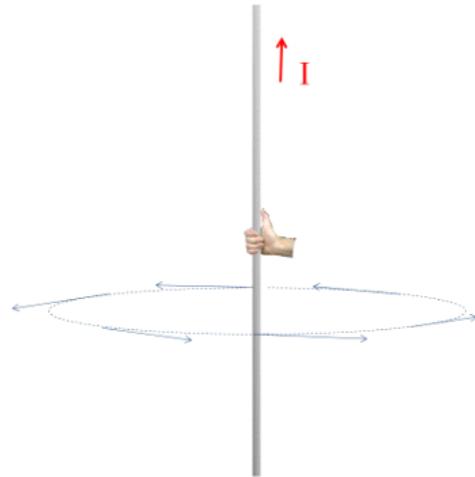


Question 223.38.2

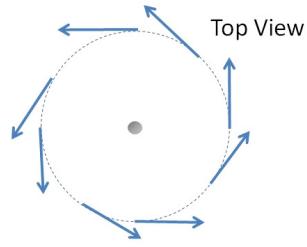
Although this is true, it takes some interpretation Let's take some time to see what it means. Let's redraw the figure.



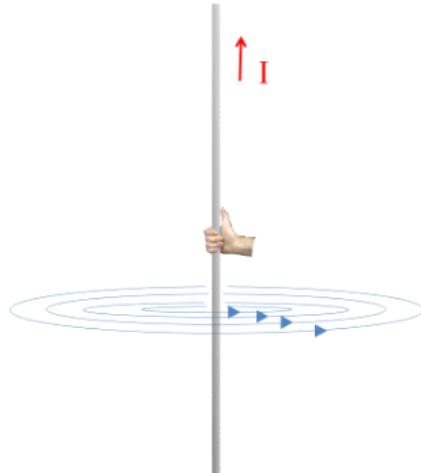
Now that we have a new figure, let's reconsider what our right hand rule means. What we mean is that the magnetic field is constant in magnitude around a circle, and that the direction of the field is tangent to the circle, with the arrow pointing in the direction your fingers go with the right-hand-rule.



This is easier to see in a top-down view.

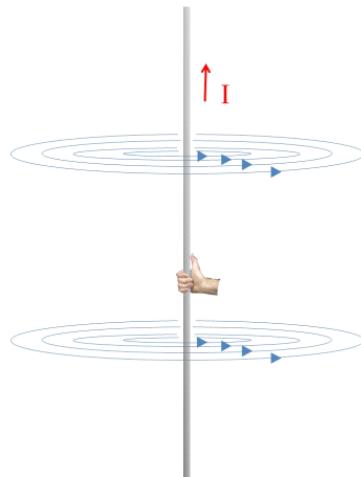


But in the first figure we only drew the field around one circle. By using symmetry, we can guess that the field magnitude must be constant around any circle. It must depend only on r , if the current is constant. So we could draw constant field lines at any distance, r , away from the wire.

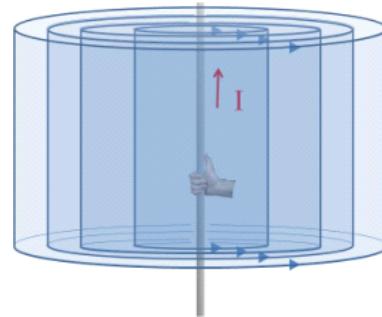


Question 223.38.3

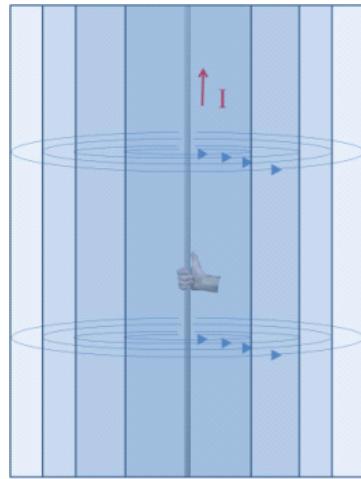
But again, this figure is not so good, because the entire wire makes a field that has a constant value for B at a distance r away. So we could also draw the field above our hand.



Maybe a better way to draw this field would be a set of concentric cylinders. Along the surface of the cylinder (but not the end caps) the field will be constant.

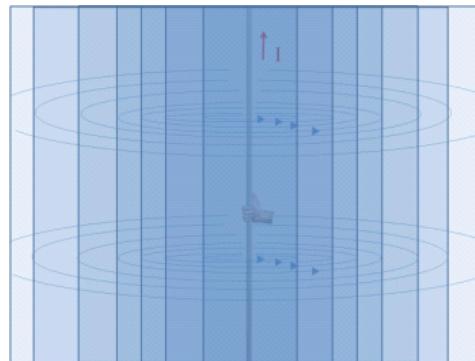


Of course, if our wire is infinitely long, the cylinders will be infinitely long too...



Question 223.38.4

And the field does not stop after a few cylinders, it reaches $B = 0$ only when $r = \infty$. So the field fills all of space.



This is a more accurate way to draw the magnetic field due to a long straight wire, but it takes a long time to draw such a diagram, so usually we will just draw one circle, and you will have to mentally fill in the other circles and the concentric cylinders that they represent.

To use the right hand rule, remember to place your thumb in the current direction. Then the field direction is given tangent to the circle and pointing in our finger direction.

Making the field—moving charges

But how does a current in a wire make a magnetic field?

The secret is to look at the individual charges that are moving. When early scientists caused individual charges to move, they found they created magnetic fields. The experimental results gave a relationship for the strength of this field

$$B = \frac{\mu_0}{4\pi} \frac{qv \sin \theta}{r^2}$$

and the direction is given by the right hand rule by pointing the thumb in the direction the charges are going and using the figures to indicate the field direction as we have described above. In a sense, this is a very small current (one moving charge!). So the field should look very similar.

This relationship was found by two scientists, Biot and Savart, and it carries their name, the *Biot-Savart law*. The factor μ_o is a constant very like ϵ_o . It has a value

$$\mu_o = 4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}$$

and is called the *permeability of free space*. The unit T is called a *tesla* and is

$$\text{T} = \frac{\text{N}}{\text{A m}}$$

The charges already had an electric field before they were accelerated, but now they

have two fields, an electric and a magnetic field. We used unit vectors to write our E -field.

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r}$$

It is convenient to do the same for the magnetic case. We can remember that a vector cross product is given by

$$\vec{a} \times \vec{b} = ab \sin \theta \quad \perp \vec{a}, \perp \vec{b}$$

where the resulting vector is perpendicular to both \vec{a} and \vec{b} . Thinking about this for a while allows us to realize this is just what we want for the magnetic field. If the velocity of the charges is up (say, in the \hat{z} direction) then we can use our right hand rule to realize we need a vector perpendicular to both \hat{z} and \hat{r} . This is given by

$$\hat{z} \times \hat{r}$$

which is always tangent to the circle indicated by our fingers. Since v is in the z direction we can use

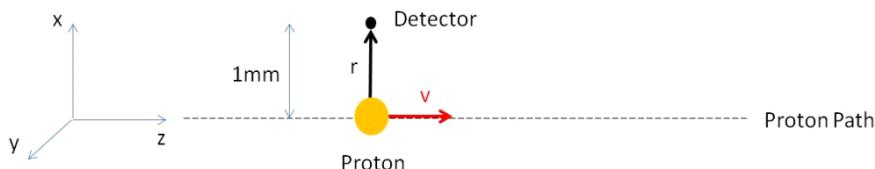
$$\vec{v} \times \hat{r} = v \sin \theta \quad \perp \vec{v}, \perp \hat{r}$$

to write the Biot-Savart law as

$$\vec{B} = \frac{\mu_0}{4\pi} \frac{q \vec{v} \times \hat{r}}{r^2}$$

We should do a problem to see how this works.

Suppose we accelerate a proton and send it in the z -direction to a speed of 1.0×10^7 m/s. Let's further suppose we have a magnetic field detector placed 1 mm from the path of the proton. What field would it measure?



We know

$$\vec{B} = \frac{\mu_0}{4\pi} \frac{q \vec{v} \times \hat{r}}{r^2}$$

and by symmetry we know that v is perpendicular to \hat{r} just as the proton passes the detector. So, using the right hand rule for cross products, we put our hand in the v -direction and bend our fingers into the r -direction. Then our thumb shows the resulting direction. In this case it is in the positive y -direction, or out of the page. The

magnitude would be

$$\begin{aligned}\vec{B} &= \frac{4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}}{4\pi} \frac{(1.6 \times 10^{-19} \text{ C}) (1.0 \times 10^7 \text{ m/s})}{(0.001 \text{ m})^2} \hat{y} \\ &= 1.6 \times 10^{-13} \text{ T} \hat{y}\end{aligned}$$

39 Current loops

Fundamental Concepts

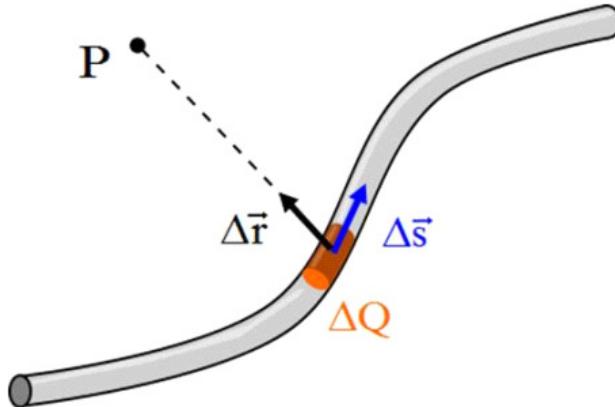
- The magnetic field due to a current in a wire is given by the integral form of the Biot-Savart law $\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{d\vec{s} \times \hat{r}}{r^2}$
- The magnetic field magnitude of a long straight wire with a current is given by $B = \frac{\mu_0 I}{2\pi a}$ with the direction given by the right hand rule we learned last time.
- The field due to a magnetic dipole is $\vec{B} \approx \frac{\mu_0}{4\pi} \frac{2\vec{\mu}}{r^3} \hat{r}$ where $\vec{\mu}$ is the magnetic dipole moment $\mu = IA$ with the direction from south to north pole.

Magnetic field of a current

Last lecture, we learned the Biot-Savart law

$$\vec{B} = \frac{\mu_0}{4\pi} \frac{q \vec{v} \times \hat{r}}{r^2}$$

now let's consider our q to be part of a current in a wire. A small amount of current moves along the wire. Let's call this small amount of charge ΔQ .



This small amount of charge will make a magnetic field, but it will be only a small part of the total field, because ΔQ is only a small part of the total amount of charge flowing

in the wire. That part of the field made by ΔQ is

$$\Delta \vec{B} = \frac{\mu_0}{4\pi} \frac{\Delta Q \vec{v} \times \hat{r}}{r^2}$$

Let's look at $\Delta Q \vec{v}$. We can rewrite this as

$$\begin{aligned}\Delta Q \vec{v} &= \Delta Q \frac{\Delta \vec{s}}{\Delta t} \\ &= \frac{\Delta Q}{\Delta t} \Delta \vec{s} \\ &= I \Delta \vec{s}\end{aligned}$$

then our small amount of field is given by

$$\Delta \vec{B} = \frac{\mu_0}{4\pi} \frac{I \Delta \vec{s} \times \hat{r}}{r^2}$$

as usual, where there is a Δ , we can predict that we can take a limit and end up with a d

$$d \vec{B} = \frac{\mu_0}{4\pi} \frac{I d \vec{s} \times \hat{r}}{r^2}$$

Question 223.39.1

Question 223.39.2

Question 223.39.3

Question 223.39.4

Question 223.39.5

Some things to note about this result

1. The vector $d \vec{B}$ is perpendicular to $d \vec{s}$ and to the unit vector \hat{r} directed from $d \vec{s}$ to some point P .
2. The magnitude of $d \vec{B}$ is inversely proportional to r^2
3. The magnitude of $d \vec{B}$ is proportional to the current
4. The magnitude of $d \vec{B}$ is proportional to the length of $d \vec{s}$
5. The magnitude of $d \vec{B}$ is proportional to $\sin \theta$ where θ is the angle between $d \vec{s}$ and \hat{r}

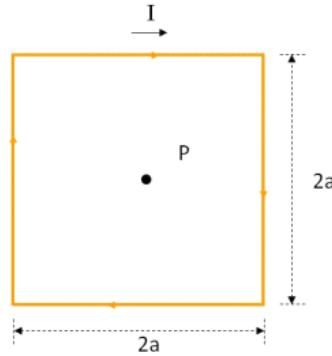
Where there is $d \vec{B}$ we will surely integrate. The field $d \vec{B}$ is due to just a small part of the wire $d \vec{s}$. We would like the field due to all of the wire. So we take

$$\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{d \vec{s} \times \hat{r}}{r^2}$$

This is a case where the equation actually is as hard to deal with as it looks. The integration over a cross product is tricky. Let's do an example.

The field due to a square current loop

Suppose we have a square current loop. Of course there would have to be a battery or some potential source in the loop to make the current, but we will just draw the loop with a current as shown. The current must be the same in all parts of the loop.



Let's find the field in the center of the loop at point \$P\$.

I will break up the integration into four parts, one for each side of the loop. For each part, we will need to find \$\vec{ds} \times \hat{r}\$ and \$r\$ to find the field using

$$\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{d\vec{s} \times \hat{r}}{r^2}$$

This is very like what we did to find electric fields. For electric fields we had to find \$dq\$, \$\hat{r}\$, and \$r\$ and we integrated using

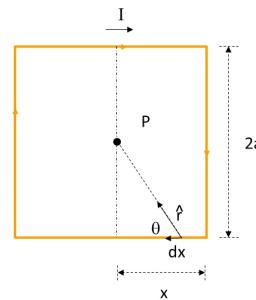
$$\vec{E} = \frac{1}{4\pi\epsilon_0} \int \frac{dq}{r^2} \hat{r}$$

Now we need \$\vec{ds} \times \hat{r}\$ and \$r\$. For electric fields, we needed to deal with the vector \$\hat{r}\$.

Now we need to deal with a cross product, \$\vec{ds} \times \hat{r}\$, involving \$\hat{r}\$. For the bottom part of our loop \$\vec{ds} \times \hat{r}\$ is just

$$\begin{aligned} d\vec{s} \times \hat{r} &= -ds \sin \theta \hat{k} \\ &= -dx \sin \theta \hat{k} \end{aligned}$$

where \$+\hat{k}\$ is out of the page. We can see this in the figure



So our field from the bottom wire is

$$\begin{aligned}\vec{B}_b &= \frac{\mu_0 I}{4\pi} \int \frac{d\vec{s} \times \hat{r}}{r^2} \\ &= \frac{\mu_0 I}{4\pi} \int \frac{-dx \sin \theta \hat{k}}{r^2}\end{aligned}$$

Next we need to find r . We would like to not have more than one variable. So it would be good to try to pick x or θ and to put everything in terms of that one variable. Let's try θ . From trigonometry we realize

$$\sin \theta = \frac{a}{r}$$

on the right side of the wire, and

$$\sin(\pi - \theta) = \sin \theta = \frac{a}{r}$$

on the left side, So all along the bottom wire r is given by

$$r = \frac{a}{\sin \theta}$$

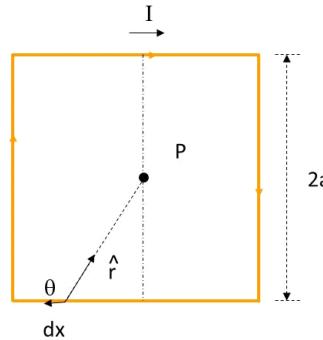
Then our field equation for the bottom wire becomes

$$\vec{B}_b = \frac{\mu_0 I}{4\pi} \int \frac{-(dx) \sin \theta \hat{k}}{\left(\frac{a}{\sin \theta}\right)^2}$$

but now we have an integration over dx and our function is in terms of θ which depends on x . We should try to fix this. Let's find dx in terms of $d\theta$. We can pick $x = 0$ to be the middle of the wire. Then

$$\tan \theta = \frac{a}{x}$$

on the right and



$$\tan(\pi - \theta) = -\tan \theta = \frac{a}{x}$$

on the left. Since on the left x is negative, this makes sense. So we have either

$$x = \frac{a}{\tan \theta}$$

or

$$x = -\frac{a}{\tan \theta}$$

depending on which size of the dotted line we are on. We could write these as

$$x = \pm \frac{a}{\tan \theta} = \pm \frac{a \cos \theta}{\sin \theta}$$

for both cases. We really want dx and moreover we want it as a magnitude (we deal with the direction in the cross product). So we can take a derivative and then take the magnitude (absolute value).

$$\frac{dx}{d\theta} = \frac{\sin \theta (-a \sin \theta) - a \cos \theta \cos \theta}{\sin^2 \theta} = \frac{-a}{\sin^2 \theta}$$

This derivative was not obvious! We had to use the quotient rule. But once we have found it we can rewrite dx as

$$dx = \left| \frac{-a}{\sin^2 \theta} d\theta \right|$$

(now with the absolute value inserted) and since neither a nor $\sin^2 \theta$ can be negative we can just write this as

$$dx = \frac{a}{\sin^2 \theta} d\theta$$

When we put this in our integral equation for the bottom wire we have

$$\vec{B}_b = \frac{\mu_o I}{4\pi} \int \frac{-\left(\frac{a}{\sin^2 \theta}\right) d\theta \sin \theta \hat{k}}{\left(\frac{a}{\sin \theta}\right)^2}$$

which we should simplify before we try to integrate.

$$\begin{aligned} \vec{B}_b &= \frac{\mu_o I}{4\pi} \int \frac{-\sin \theta d\theta \hat{k}}{a} \\ &= -\frac{\mu_o I}{4\pi a} \hat{k} \int \sin \theta d\theta \end{aligned}$$

which is really not too bad considering the integral we had at the start of this problem. When we get to the corner of the left hand side $\theta = \frac{3\pi}{4}$ and when we start on the right hand side $\theta = \frac{\pi}{4}$ and along the bottom wire θ will be somewhere in between $\frac{\pi}{4}$ and $\frac{3\pi}{4}$. Then $\frac{\pi}{4}$ and $\frac{3\pi}{4}$ are our limits of integration. We can perform this integral

$$\begin{aligned} \vec{B}_b &= -\frac{\mu_o I}{4\pi a} \hat{k} \int_{\frac{\pi}{4}}^{\frac{3\pi}{4}} \sin \theta d\theta \\ &= -\frac{\mu_o I}{4\pi a} \hat{k} [-\cos \theta]_{\frac{\pi}{4}}^{\frac{3\pi}{4}} \\ &= -\frac{\mu_o I}{4\pi a} \hat{k} \left(\frac{\sqrt{2}}{2} - \left(-\frac{\sqrt{2}}{2} \right) \right) \\ &= -\frac{\mu_o I \sqrt{2}}{4\pi a} \hat{k} \end{aligned}$$

This was just for the bottom of the loop. Now let's look at the top of the loop. There is finally some good news. The math will all be the same except for the directions. We had better work out $d\vec{s} \times \hat{r}$ to see how different it is.

Now the $d\vec{s}$ is to the right and \hat{r} is downward so

$$d\vec{s} \times \hat{r} = -dx \sin \theta \hat{k}$$

But this is just as before. So even this is the same! The integral across the top wire will have exactly the same result as the integral across the bottom wire. We can just multiply our previous result by two.

How about the sides? Again we get the same $d\vec{s} \times \hat{r}$ direction and all the rest is the same, so our total field is

$$\vec{B} = 4\vec{B}_b = -\frac{\mu_o I \sqrt{2}}{\pi a} \hat{k}$$

This was a long hard, messy problem. But current loops are important! Every electric circuit is a current loop. Does this mean that every circuit is making a magnetic field?

Question 223.39.6

The answer is yes! As you might guess, this can have a profound effect on circuit design. If your circuit is very sensitive, adding extra fields (and therefore extra forces on the charges) can be disastrous causing the design to fail. There is some concern about “electronic noise” and possible effects on the body (cataracts are one side effect that is well known). And of course, as the circuit changes its current, the field it creates changes. This can create the opportunity for espionage. The field exists far away from the circuit. A savvy spy can determine what your circuit is doing by watching the field change!

Long Straight wires

In our last example, we found that the magnitude of the field due to a wire is

$$B = \left| -\frac{\mu_o I}{4\pi a} \int \sin \theta d\theta \right|$$

Of course, we would like to relate this to our standard charge configuration, in this case an infinite line of (now moving) charge. If the wire is infinitely long, then the limits of integration are just from $\theta = 0$ to $\theta = \pi$

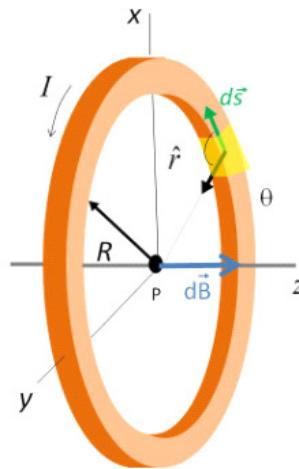
$$\begin{aligned} B &= \left| -\frac{\mu_o I}{4\pi a} \int_0^\pi \sin \theta d\theta \right| \\ &= \left| -\frac{\mu_o I}{4\pi a} (-\cos \theta) \Big|_0^\pi \right| \\ &= \frac{\mu_o I}{2\pi a} \end{aligned}$$

This is an important result. We can add a new geometry to our list of special cases, a long straight wire that is carrying a current I . The direction of the magnetic field, we already know, is given by our right-hand-rule. Of course, if our wire is not infinitely

long, we now know how to find the actual field. It is all a matter of finding the right limits of integration.

Question 223.39.7

Magnetic dipoles



As a second example, let's find the magnetic field due to a round loop at the center of the loop. We start again with

$$\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{d\vec{s} \times \hat{r}}{r^2}$$

We need to find $d\vec{s} \times \hat{r}$ and r , to do the integration. Our steps are:

1. Find an expression for $d\vec{s} \times \hat{r}$
2. Find an expression for r
3. Assemble the integral, including limits of integration
4. Solve the integral.

Let's start with the first step. As we go around the loop $d\vec{s}$ and \hat{r} will be perpendicular to each other, so

$$ds \times \hat{r} = ds \hat{k}$$

For the second step, we realize that r is just the radius of the loop, R . Then the

integration is quite easy (much easier to set up than the last case!)

$$B = \frac{\mu_0 I}{4\pi} \int \frac{ds}{R^2} \hat{k}$$

The limits of integration will be 0 to $2\pi R$. We can perform this integral

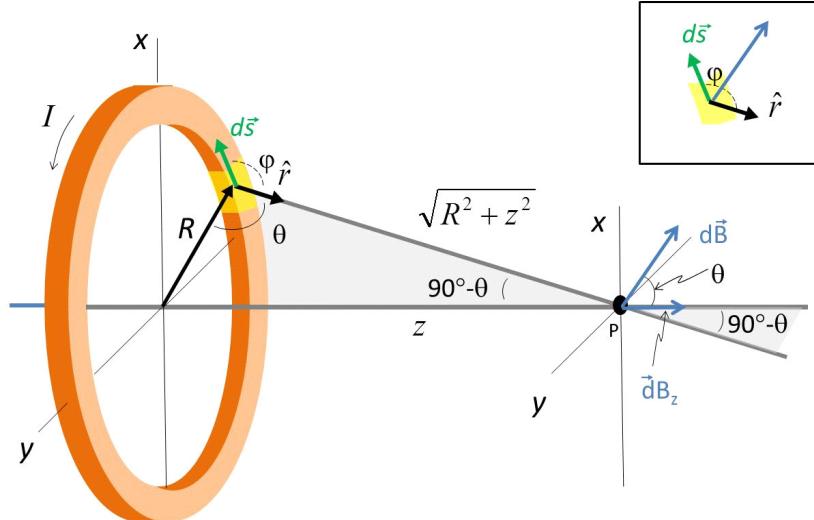
$$\begin{aligned} B &= \frac{\mu_0 I}{4\pi} \int_0^{2\pi R} \frac{ds}{R^2} \hat{k} \\ &= \frac{\mu_0 I}{4\pi} \frac{2\pi R}{R^2} \hat{k} \end{aligned}$$

so

$$B = \frac{\mu_0 I}{2R} \hat{k} \quad \text{loop}$$

The field is perpendicular to the plane of the loop, which agrees with our square loop problem.

Let's extend this problem to a point along the axis a distance z away from the loop.



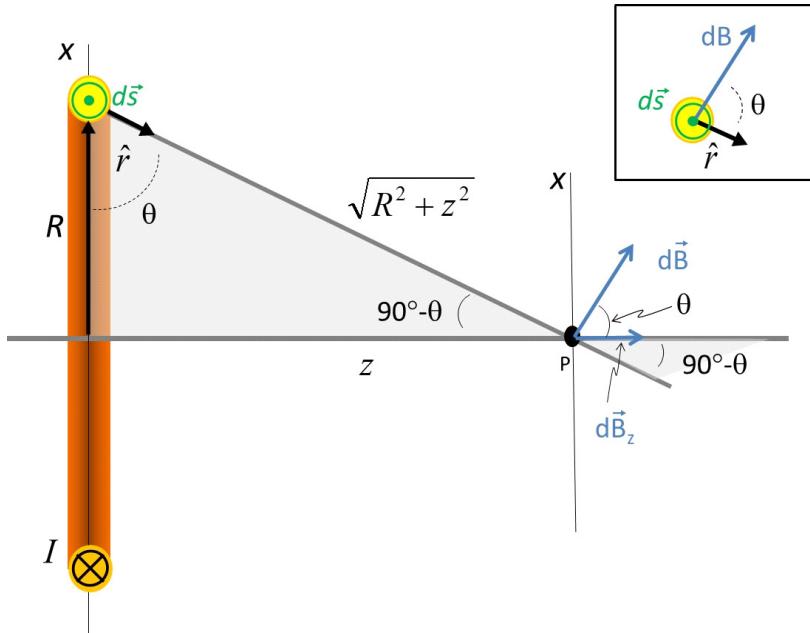
We need to go back to our basic equation again.

$$\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{d\vec{s} \times \hat{r}}{r^2}$$

Starting with step 1, we realize that, in general, our value of $d\vec{s} \times \hat{r}$ is

$$d\vec{s} \times \hat{r} = ds \sin \phi$$

where ϕ is the angle between $d\vec{s}$ and \hat{r} . We can see that for this case ϕ will still be 90° .



We have tipped \hat{r} toward our point P , but tipping \hat{r} from pointing to the center of the hoop to pointing to a point on the axis just rotated \hat{r} about part of the hoop. We still have $\phi = 90^\circ$. So

$$|d\vec{s} \times \hat{r}| = ds$$

with a direction shown in the figure. We have used symmetry to argue that we can take just x or y -components in the past because all the others clearly canceled out. We can also do that again here. Using symmetry we see that only the z -component of the magnetic field will survive. So we can take the projection onto the z -axis.

$$\vec{B} = \frac{\mu_0 I}{4\pi} \int \frac{ds}{r^2} \cos \theta \hat{k}$$

We know how to deal with such a situation, since we have done this before. From the diagram we can see that

$$\cos \theta = \frac{R}{\sqrt{R^2 + z^2}}$$

And our value of r is now more complicated, but in a way we recognize

$$r = \sqrt{R^2 + z^2}$$

so our field becomes

$$\vec{B} = \frac{\mu_0 I}{4\pi} \hat{k} \int \frac{R ds}{(R^2 + z^2)^{\frac{3}{2}}}$$

Fortunately this integral is also not too hard to do. Let's take out all the terms that don't change with ds

$$\vec{B} = \frac{\mu_0 I R}{4\pi (R^2 + z^2)^{\frac{3}{2}}} \hat{k} \int_0^{2\pi R} ds$$

The limits of integration are 0 to $2\pi R$, the circumference of the circle

$$\begin{aligned}\vec{B} &= \frac{\mu_0 I R 2\pi R}{4\pi (R^2 + z^2)^{\frac{3}{2}}} \hat{k} \\ &= \frac{\mu_0 I R^2}{2(R^2 + z^2)^{\frac{3}{2}}} \hat{k}\end{aligned}$$

Let's take some limiting cases to see if this makes sense. Suppose $z = 0$, then

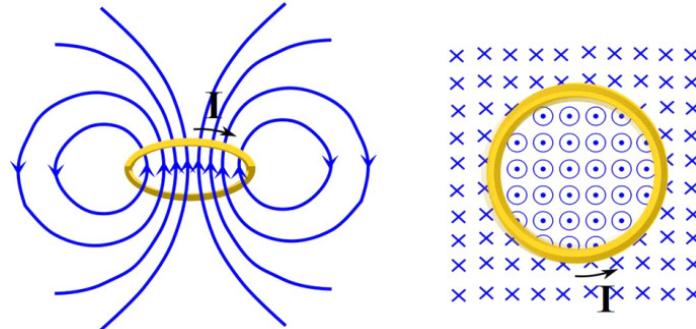
$$\begin{aligned}\vec{B} &= \frac{\mu_0 I R^2}{2(R^2 + 0)^{\frac{3}{2}}} \hat{k} \\ &= \frac{\mu_0 I R^2}{2R^3} \hat{k} \\ &= \frac{\mu_0 I}{2R} \hat{k}\end{aligned}$$

which is what we got before for the field at the center of the loop. That is comforting.

Now suppose that $z \gg R$. In that case, we can ignore the R^2 in the denominator.

$$\begin{aligned}\vec{B} &\approx \frac{\mu_0 I R^2}{2(z^2)^{\frac{3}{2}}} \hat{k} \\ &= \frac{\mu_0 I R^2}{2z^3} \hat{k}\end{aligned}$$

We have just done this for on-axis positions because the math is easy there. But we could find the field at other locations. The result looks something like this.



The figure on the left was taken from the pattern in iron filings that was created by an actual current loop field. The figure to the right is a top down look. We will use the symbol \odot to mean “coming out of the page at you” and the symbol \times “going into the page away from you.” Imagine these as parts of an arrow. The dot in the circle is the arrow tip coming at you, and the cross is the fletching going away from you. Notice that the field is up through the loop, and down on the outside.

As we generalize our solution for the magnetic field far from the loop we have

$$\vec{B} \approx \frac{\mu_0 I R^2}{2r^3} \hat{k}$$

This looks a lot like the electric field from a dipole

$$\vec{E} = \frac{2}{4\pi\epsilon_0} \frac{\vec{p}}{r^3}$$

which gives us an idea. We have a dipole moment for the electric dipole. This magnetic field has the same basic form as the electric dipole. We can rewrite our field as

$$\begin{aligned}\vec{B} &\approx \frac{\mu_0 I (\pi R^2)}{2(\pi) r^3} \hat{i} \\ &= \frac{\mu_0 (2) I (A)}{(2) 2(\pi) r^3} \hat{i} \\ &= \frac{\mu_0 2IA}{4\pi r^3} \hat{i}\end{aligned}$$

where $A = \pi R^2$ is the area of the loop.

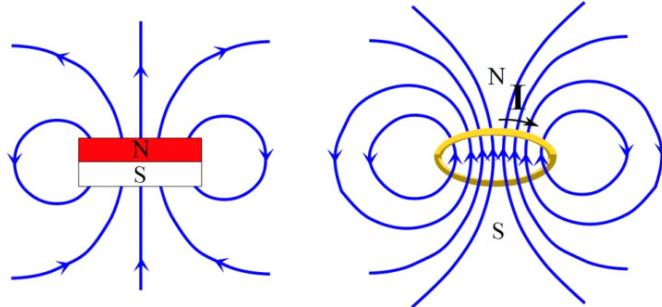
The electric dipole moment is the charge multiplied by the charge separation

$$p = qa$$

we have something like that in our magnetic field, The terms IA describe the amount of charge and the geometry of the charges. We will call these terms together the *magnetic dipole moment*

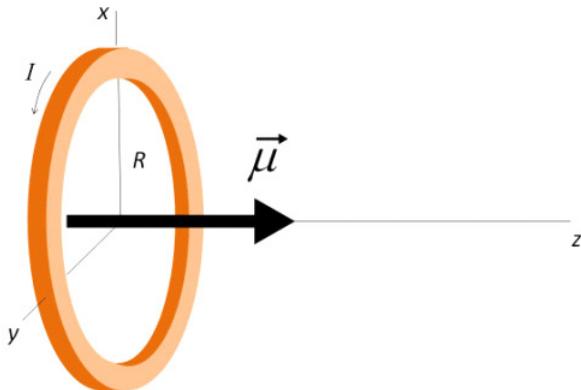
$$\mu = IA$$

and give them a direction so that μ is a vector. The direction will be from south to north pole



where we can find the south and north poles by comparison to the field of a bar magnet.

$$\vec{\mu} = IA \quad \text{from South to North}$$



This is a way to characterize an entire current loop.

As we get farther from a loop, the exact shape of the loop becomes less important. So as long as r is much larger than R , we can write

$$\vec{B} \approx \frac{\mu_0}{4\pi} \frac{2\vec{\mu}}{r^3} \hat{k}$$

for any shaped current loop.

The integral form of the Biot-Savart law is very powerful. We can use computers to calculate the field due to any type of current configuration. But by hand there are only a few cases we can do because the integration becomes difficult. With electrostatics, we found ways to use geometry to eliminate or at least make the integration simpler. We will do the same thing for magnetostatics starting with the next lecture. Our goal will be to use geometry to avoid using Biot-Savart when we can.

Basic Equations

40 Ampere's law, and Forces on Charges

Fundamental Concepts

- The magnetic field can be found more simply for symmetric currents using Ampere's law $\oint \vec{B} \cdot d\vec{s} = \mu_0 I_{through}$
- The force due to the magnetic field on a charge, q , is given by $\vec{F} = q \vec{v} \times \vec{B}$

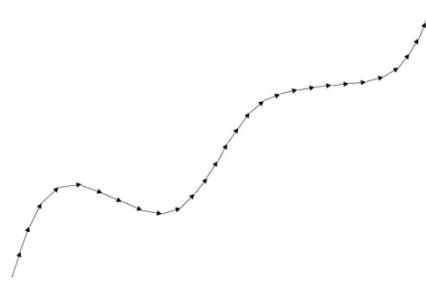
Ampere's Law

The Biot-Savart law is a powerful technique for finding a magnetic field, but it is more powerful numerically than in closed-form problems. We can only find exact solutions to a few problems with special symmetry. Since problems we can do by hand require special symmetry anyway, we would like to use symmetry as much as possible to remove the need for difficult integration.

We saw this situation before with electrostatics. We did some integration to find fields from charge distributions, but then we learned Gauss' law, and that was easier because it turned hard integration problems into relatively easy ones. This still required special symmetry, but when it worked, it was a fantastic time saver. For non-symmetric problems, there is always the integration method, and a computer.

Likewise, for magnetostatics there is an easier method. To see how it works, let's review some math.

In the figure there is a line, divided up into many little segments.



We can find the length of the line by adding up all the little segment lengths

$$L = \sum_i \Delta s_i$$

Integration would make this task less tedious

$$L = \int ds$$

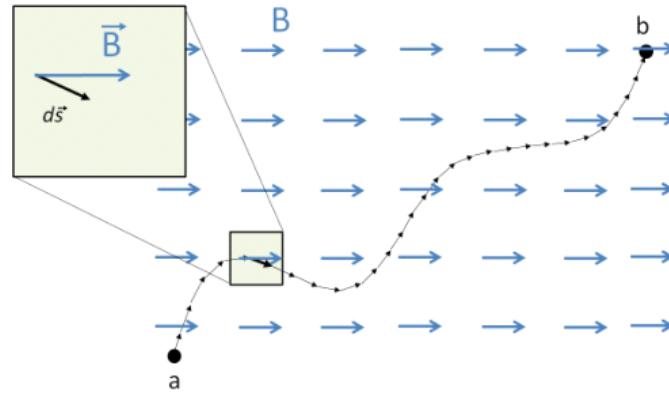
This is called a line integral. Our new method of finding magnetic fields will involve line integrals. The calculation of the length is too simple, however. We will have to integrate some quantity along the line. For example, we could envision integrating the amount of energy lost when pushing a box along a path. The integral would give the total energy loss. The amount of energy lost would depend on the specific path. Thus a line integral

$$W = \int \vec{F} \cdot d\vec{s}$$

would be useful to find the total amount of work. Each small line segment would give a differential amount of work

$$dW = \vec{F} \cdot d\vec{s}$$

and we use the integral to add up the contribution to the work for each segment of size ds along the path. Notice the dot product. We need the dot product because only the component of the force in the direction the box is going adds to the total work done.



We wish to do a similar thing for our magnetic field. We wish to integrate the magnetic

field along a path. The integral would look like this

$$\int_a^b \vec{B} \cdot d\vec{s}$$

This may not look like an improvement over integrating using the Biot-Savart law, but our goal will be to use symmetry to make this integral very easy. The key is in the dot product. We want only the component of the magnetic field that is in the $d\vec{s}_i$ direction. There are two special cases.

If the field is perpendicular to the $d\vec{s}_i$ direction, then

$$\int_a^b \vec{B} \cdot d\vec{s} = 0$$

because $\vec{B} \cdot d\vec{s}_i = 0$ for this case

If the field is in the same direction as $d\vec{s}_i$, then $\vec{B} \cdot d\vec{s}_i = Bds$ and

$$\int_a^b \vec{B} \cdot d\vec{s} = \int_a^b Bds$$

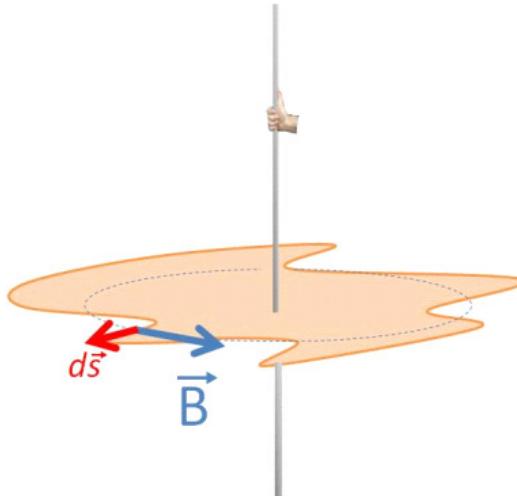
Further if we can make is so that B is constant and everywhere tangent to the path, then

$$\begin{aligned} \int_a^b \vec{B} \cdot d\vec{s} &= \int_a^b Bds \\ &= B \int_a^b ds \\ &= BL \end{aligned}$$

This process should look familiar. We used similar arguments to make the integral $\int \vec{E} \cdot d\vec{A}$ easy for Gauss' law.

With Gaussian surfaces, we found we could imagine any surface we wanted. In a similar way, for our line integral we can pick any path we want. if we can make B constant and everywhere tangent to the path, then, the integral will be easy. It is important to realize that we get to make up our path. There may be some physical thing along the path, but there is no need for there to be. The paths we will use are imaginary.

Usually we will want our path to be around a closed loop. Let's take the case of a long straight current-carrying wire. We know the field shape for this. We can see that if we take a crazy path around the wire, that $\vec{B} \cdot \Delta\vec{s}_i$ will give us the projection of \vec{B} onto the $\Delta\vec{s}_i$ direction for each part of the path.



We get

$$\sum_i B_{\parallel} \Delta s$$

where B_{\parallel} is the component of B that is parallel to the Δs direction. In integral form this is

$$\int B_{\parallel} ds$$

The strange shape I drew is not very convenient. This is neither the case where $\vec{B} \cdot d\vec{s}_i = 0$ nor where $\vec{B} \cdot d\vec{s}_i = Bds$. But if we think for a moment, I do know a shape where $\vec{B} \cdot d\vec{s}_i = Bds$. If we choose a circle, then from symmetry B will be constant, and it will be in the same direction as ds so $\vec{B} \cdot d\vec{s}_i = Bds$. From our last lecture we even know what the field should be for a long straight wire.

$$B = \frac{\mu_0 I}{2\pi r}$$

Let's see if we can use this to form a new general approach. Since B is constant around the loop (because r is constant around the loop), we can write our line integral as

$$\begin{aligned} \int \vec{B} \cdot d\vec{s} &= B2\pi r \\ &= \frac{\mu_0 I}{2\pi r} 2\pi r \\ &= \mu_0 I \end{aligned}$$

This is an amazingly simple result. We integrated the magnetic field around an imaginary loop path, and got that the result is proportional to the current in the wire. This reminds us of Gauss' law where we integrated the electric field around a surface and got that the result is proportional to the amount of charge inside the surface.

$$\int \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\epsilon_0}$$

Let's review. Why did I pick a circle as my imaginary path? Because it made my math easy! I don't want to do hard math to compute the field, so I tried to find a path over which the math was as easy as possible. Since the path is imaginary, I can choose any path I want, so I chose a simple one. I want a path where $\vec{B} \cdot d\vec{s}_i = 0$ or where $\vec{B} \cdot d\vec{s}_i = Bds$. This is very like picking Gaussian surfaces for Gauss' law. If I chose a harder path I would get the same answer, but it would take more effort. I found the result of my integral $\int \vec{B} \cdot d\vec{s}$ to be just $\mu_o I$.

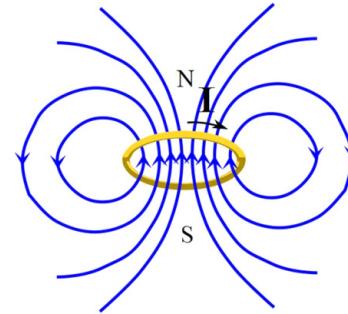
We had to integrate around a closed path, so I will change the integral sign to indicate that we integrated over a closed path.

$$\oint \vec{B} \cdot d\vec{s} = \mu_o I_{through} \quad (40.1)$$

and only the current that went through the imaginary surface contributed to the field, so we can mark the current as being the current that goes through our imaginary closed path.

This process was first discovered by Ampere, so it is known as Ampere's law.

Let's use Ampere's law to do another problem. Suppose I have a coil of wire. This coil is effectively a stack of current rings. We know the field from a single ring.



$$B = \frac{\mu_o I}{2R}$$

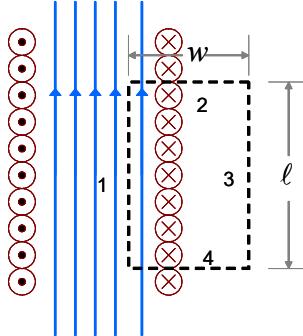
But what would the field be that is generated by having a current flow through the coil?

Well, looking at the single loop picture, we see that the direction of the field due to a loop is right through the middle of the loop. I think it is reasonable to believe that if

I place another loop on top of the one pictured, that the fields would add, making a stronger field down the middle. This is just what happens. So I could write our loop field equation as

$$B = N \frac{\mu_o I}{2r}$$

where N is the number of loops I make. It is customary in electronics to define n as the number of loops per unit length (sort of like the linear mass density we defined in waves on strings, only now it is linear loop density). Suppose I take a lot of loops! In the picture I have drawn the loops like a cross section of a spring. But now the loops are not all at the same location. So we would guess that our field will be different than just N times the field due to one loop. We can use Ampere's law to find this field?



Consider current is coming out at us on the LHS and is going back into the wires on the RHS. Remember our goal is to use Ampere's law

$$\oint \vec{B} \cdot d\vec{s} = \mu_o I_{\text{through}}$$

to find the field. Let's imagine a rectangular shaped *Ampelian* loop shown as a dotted black line. Note that like Gaussian surfaces, this is an imaginary loop. Nothing is really there along the loop. Let's look at the integral by breaking it into four pieces,

$$\int_1 \vec{B} \cdot d\vec{s} + \int_2 \vec{B} \cdot d\vec{s} + \int_3 \vec{B} \cdot d\vec{s} + \int_4 \vec{B} \cdot d\vec{s} = \mu_o I_{\text{through}}$$

one for each side of the loop. If I have chosen my loop carefully, then $\vec{B} \cdot d\vec{s}_i$ will either be $\vec{B} \cdot d\vec{s}_i = 0$ or $\vec{B} \cdot d\vec{s}_i = B ds$. Let's start with side 2. We want to consider

$$\vec{B} \cdot d\vec{s}_2$$

We see that for our side 2 the field is perpendicular to $d\vec{s}_2$ So

$$\mathbf{B} \cdot d\ell_2 = 0$$

This is great! I can integrate 0

$$\int 0 = 0$$

The same reasoning applies to

$$\vec{B} \cdot d\vec{s}_4 = 0$$

From our picture we can see that there is very little field outside of our coil of loops. So B_3 is very small, so $\vec{B} \cdot d\vec{s}_3 \approx 0$. It is not exactly zero, but it is small enough that I will call it negligible for this problem. For an infinite coil, this would be exactly true (but infinite coils are hard to build).

That leaves path 1. There the B -field is in the same direction as $d\vec{s}_1$ so

$$\vec{B} \cdot d\vec{s}_1 = B ds_1$$

Again this is great! B is fairly uniform along the coil. Let's say it is close enough to be considered constant. Then the integral is easy over side 1

$$\int B ds_1 = B\ell$$

We have performed the integral!

$$\begin{aligned} \oint \vec{B} \cdot d\vec{s} &= \int_1 \vec{B} \cdot d\vec{s} + \int_2 \vec{B} \cdot d\vec{s} + \int_3 \vec{B} \cdot d\vec{s} + \int_4 \vec{B} \cdot d\vec{s} \\ &= B\ell + 0 + 0 + 0 \\ &= B\ell \end{aligned}$$

Now we need to find the current in the loop. This is more tricky than it might appear. It is not just I because we have several loops that go through our loop, each on its own carrying current I and each contributing to the field. We can use a linear loop density³¹ n to find the number of loops.

$$N = n\ell$$

and the current inside the loop will be

$$I_{inside} = NI$$

Then, putting the integration all together, we have

$$\oint \vec{B} \cdot d\vec{s} = B\ell + 0 + 0 + 0 = \mu_o NI$$

or

$$B\ell = \mu_o NI$$

which gives a field of

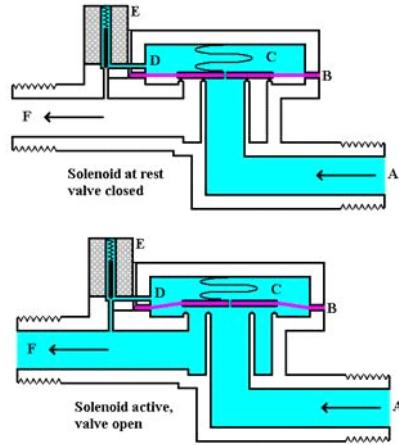
$$B = \mu_o \frac{N}{\ell} I$$

or

$$B = \mu_o n I$$

This device is so useful it has a name. It is called a *solenoid*. You may have made a coil as a kid and turned it into an electromagnet by hooking it to a battery (a source of potential difference) so that a current ran through it. In engineering solenoids are used as current controlled magnetic switches.

³¹ Physicists like densities!



Solenoid operated valve system.

There is another great thing about a solenoid. In the middle of the solenoid, the field is really nearly constant. Near the ends, there are edge effects, but in the middle we have a very uniform field. This is analogous to the nearly uniform electric field inside a capacitor. We can therefore see how to generate uniform magnetic fields and consider uniform B -fields in problems. Such a large nearly uniform magnetic field is part of the Compact Muon Solenoid (CMS) experiment at CERN.



CMS Detector at CERN. The detector is constructed of a very large solenoid to bend the path of the charge particles.

Magnetic Force on a moving charge

Now that we know how to generate a magnetic field, we can return to thinking about magnetic forces on mover charges. Our magnetic field is slightly more complicated than the electric field. We can still use a charge and the force, but now the charge is moving so we expect to have to include the velocity of the charge. We want an expression that relates B and F_{mag} in both magnitude and direction.

Our expression for the relationship between charge, velocity, field and the force comes from experiment (although now we can derive it). The experiments show that when a charged particle moves parallel to the magnetic field, there is no force! This is radically different from our E -field! Worse yet, the force seems to be perpendicular to both v and B when the angle between them is not zero! Here is our expression.

$$\mathbf{F}_B = q\mathbf{v} \times \mathbf{B} \quad (40.2)$$

where q is the mover charge and B is the magnetic field environment.

We have a device that can shoot out electrons. The electrons show up because they hit a phosphorescent screen. When we bring a magnet close to our beam of electrons, we find it moves!

But we did this with moving electrons, what happens if they are not moving? We might expect the electrons to accelerate just the same—and we would be wrong! Static charges seem to not notice the presence of the magnet at all!

We expect that, like gravity and electric charge, the force on the moving electrons must be due to a field, but this *magnetic field* does not accelerate stationary electrons. We learned before that the reason we know that there is some force on the electrons came when Oersted, a Dutch scientist experimenting with electric current, found that his compass acted strangely when it was near a wire carrying electric current. This discovery is backwards of our experiment. It implies that moving charges must effect magnets, but given Newton's third law, If moving electrons make a field that makes a force on a magnet, then we would expect a magnet will make a field that makes a force on moving charges as well!

The derivation of the magnitude of the force from the experimental data is tedious. We will just learn the results, but they are exciting enough! The magnitude of the force on a moving charge due to a constant magnetic field is

$$\vec{F}_B = q\vec{v} \times \vec{B} \quad (40.3)$$

The magnitude is given by

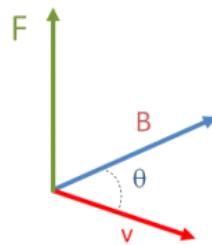
$$F = qvB \sin \theta$$

where q is charge, v is speed, and B is the magnitude of the magnetic field. We need to carefully define θ . Since we have a cross product, θ is the angle between the field direction and the velocity direction.

We can solve the equation for the magnetic field force (equation 40.3) to find the magnitude of the field

$$\frac{F}{qv \sin \theta} = B$$

But the strangeness has not ended. we need a direction of the force. And it turns out that it is perpendicular to both \vec{v} and \vec{B} as the cross product implies! We use our favorite right hand rule to help us remember.



We start with our hand pointing in the direction of $\tilde{\mathbf{v}}$. Curl your fingers in the direction of $\tilde{\mathbf{B}}$. And your fingers will point in the direction of the force. We saw this type of right hand rule before with torque, but there is one big difference. This really is the direction the charge will accelerate! Note that this works for a positive charge. If the charge is negative, then the q in

$$\vec{\mathbf{F}}_B = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$$

will be negative, and so the force will go in the other way. To keep this straight in my own mind, I still use our right hand rule, and just remember that if F is negative, it goes the opposite way of my thumb.

Right hand rule #2: We start with our hand pointing in the direction of $\tilde{\mathbf{v}}$. Curl your fingers in the direction of $\tilde{\mathbf{B}}$. And your fingers will point in the direction of the force. The magnitude of the force is given by

$$F = qvB \sin \theta \quad (40.4)$$

Motion of a charged particle in a *B*-Field

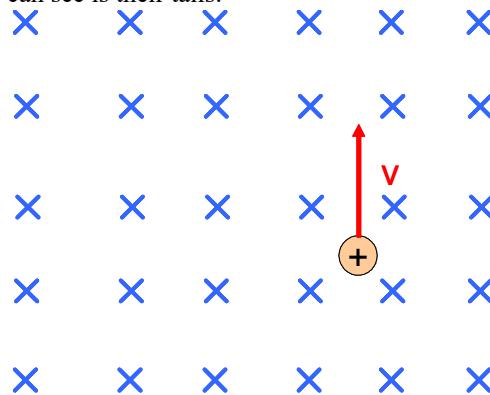
Question 223.40.1

Question 223.40.2

We refer to the magnetic field as a *B*-field for short.

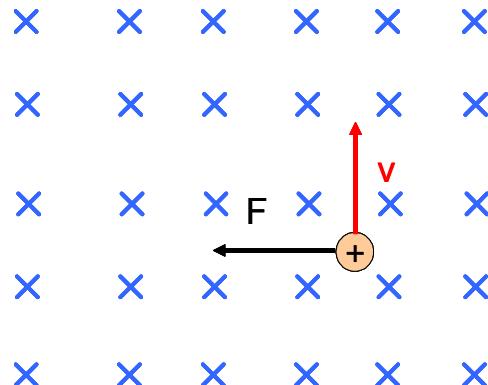
Question 223.40.3

Let's set up a constant *B*-field as shown in the figure. We draw a *B*-field as a set of vectors just like we did for electric fields. In the figure, the vectors are all pointing "into the paper" so all we can see is their tails.

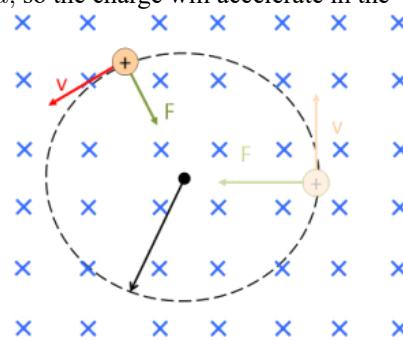


Question 223.40.6

If I have a charged particle, with velocity $\tilde{\mathbf{v}}$, what will be the motion of the particle in the field? First off, we should recall that $\tilde{\mathbf{F}}$ is in a direction perpendicular to $\tilde{\mathbf{v}}$ and $\vec{\mathbf{B}}$.. Using our right hand rule we see that it will go to the left.



Remember that $F = ma$, so the charge will accelerate in the $-x$ direction.



Now, if we allow the charged particle to move, we see that the v direction changes. This makes the direction of F change. Since v and a are always at 90° , the motion reminds us of circular motion! Let's see if we can find the radius of the circular path of the charge.

$$F = qvB \sin \theta$$

will be just

$$F = qvB$$

because θ is always 90° . Then, using Newton's second law

$$F = ma = qvB$$

and noting that the acceleration is center-seeking, and our velocity is always tangential, we can write it as a centripetal acceleration

$$a_C = \frac{v_t^2}{r}$$

Then

$$m \frac{v_t^2}{r} = qv_t B$$

$$m \frac{v_t}{r} = qB$$

We can find the radius of the circle

$$\frac{mv_t}{qB} = r$$

Could we find the angular speed?

$$\omega = \frac{v_t}{r} = \frac{qB}{m}$$

How about the period? We can take the total distance divided by the total time for a revolution

$$v_t = \frac{2\pi r}{T}$$

to find

$$T = \frac{2\pi r}{v_t}$$

and we recognize

$$\frac{1}{\omega} = \frac{r}{v_t}$$

so

$$T = \frac{2\pi}{\omega}$$

so, using our angular speed we can say

$$T = \frac{2\pi m}{qB}$$

The angular frequency ω that we found is the frequency of a type of particle accelerator called a cyclotron. This type of accelerator is used by places like CERN to start the acceleration of charged particles. The same concept is used to make the charged particles go in a circular path in the large accelerators like the LHC at CERN.



Turning magnets at CERN. This is an actual magnet, but this magnet is at ground level in the testing facility. The tunnel is a mock-up of what the actual beam tunnel looks like.

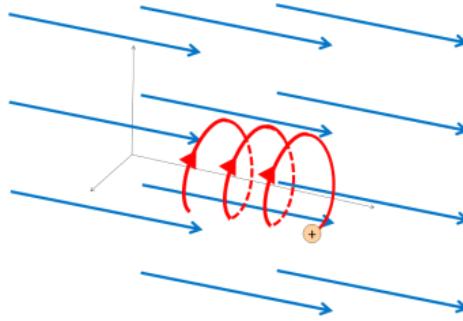
Within the detector systems, like the CMS, charged product particles can be tracked along curved paths for identification.

Question 223.40.7

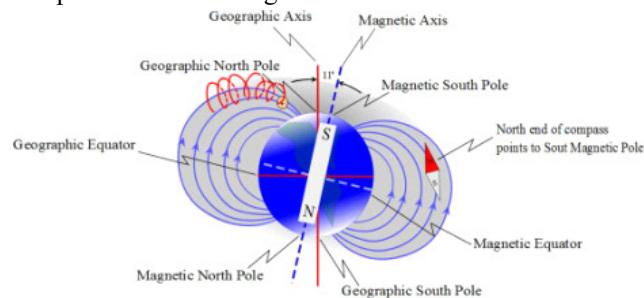
Question 223.40.8

Question 223.40.9

But it is also interesting to know that charged particles that enter a magnetic field with some initial speed will gain a circular motion as well.



An example is the charged particles from the Sun entering the Earth's magnetic field. the particles will spiral around the magnetic field lines.



As the helical motion tightens near the poles, the particles will sometimes give off patterns of light as they hit atmospheric atoms.

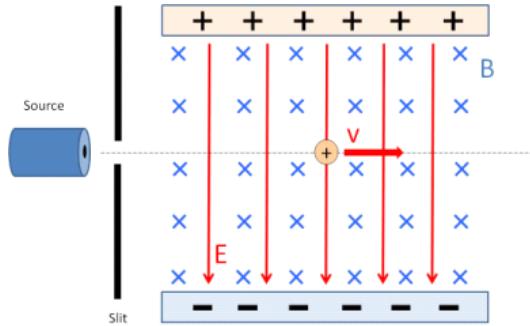


Aurora Borealis: Sand Creek Ponds Idaho 2013

The light is what we call the aurora borealis. A more high-tech use for this helical motion is the confinement of charged particles in a magnetic field for fusion reaction.

The velocity selector

Question 223.40.10



This device shows up on tests, especially finals, because it has both an electric field and a magnetic field—you test two sets of knowledge at once! So let's see how it works. Our question should be, what is the velocity of a charged particle that travels through the field without being deflected?

E-field

We remember that the force on a positively charged particle will be

$$F_E = qE$$

directed in the field direction so it is downward.

B-Field

Now we know that

$$\mathbf{F}_B = q\mathbf{v} \times \mathbf{B}$$

and we use our right hand rule to find that the direction will be upward with a magnitude of

$$\begin{aligned} F_B &= qvB \sin \theta \\ &= qvB \end{aligned}$$

So there will be no deflection (no acceleration) when the forces in the *y*-direction balance.

$$\Sigma F_y = 0 = -F_E + F_B$$

or

$$qE = qvB$$

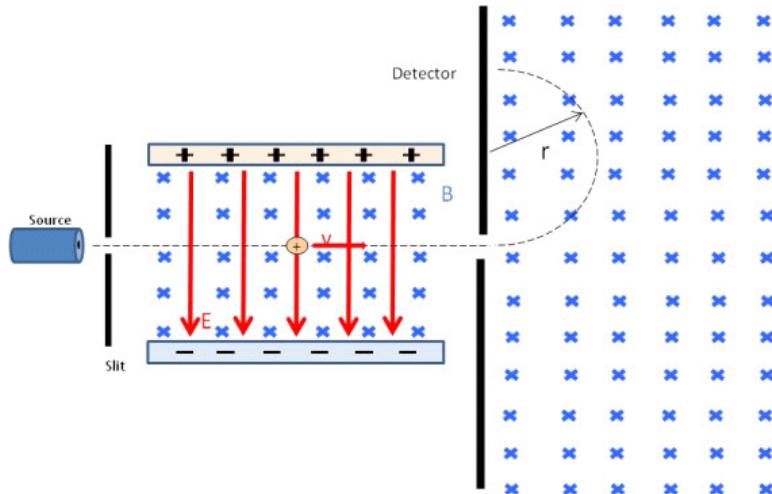
which gives

$$v = \frac{E}{B}$$

as the speed that will be "selected."

Bainbridge Mass Spectrometer

You may use a mass-spec some time in your careers. I have had samples identified by mass-spectrometers several times in my industrial career. They are very useful devices—especially when chemical identification is hard or impossible.



The Bainbridge device is one type that we can easily understand. It starts with a velocity selector which sends charged particles at a particular speed into a region of uniform magnetic field. The charged particles then follow curved paths on their way to an array of detectors. When they hit the array, their spatial location is recorded. Where they hit depends on their ratio of charge to mass. From our study of the rotational motion we found

$$r = \frac{mv}{qB_o}$$

so the charge to mass ratio is

$$\frac{q}{m} = \frac{v}{rB_o}$$

Since we know the initial velocity will be

$$v = \frac{E}{B}$$

from the velocity selector, then

$$\frac{q}{m} = \frac{E}{rBB_o}$$

One way this is often used is to separate a sample of substance, say, carbon to find the

relative amount of each isotope. The carbon atoms will all ionize to the same charge. Then the position at which they are detected depends on the mass.

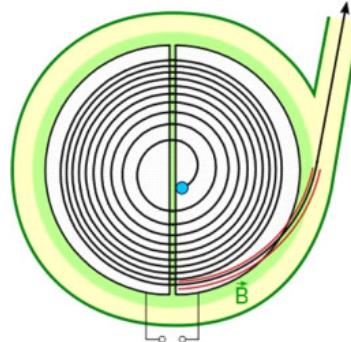
I used a mass-spec in my last industry project to identify large carbon compounds and their relative concentration in complex oil leaks. This data helped us look for possible leak detection targets so pipeline leaks could be detected before the oil was visible to the naked eye.

Classical Cyclotron

We already found the period of rotation of a charged particle in a uniform magnetic field.

$$T = \frac{2\pi m}{qB}$$

Note that this does not depend on the speed of the particle! So it will have the same travel time regardless of how fast it goes. We can use this to accelerate particles. But we add in an electric field to do the acceleration. The device is shown in the figure below



Basic Geometry of the Cyclotron. (Public Domain image courtesy KlausFoehl)

The particle starts in the center circling around in the magnetic field, but the device is divided into halves (called "Ds"). There is a gap between the Ds, and the electric field is created in the gap. One side at high potential and the other at low potential. When the particle is in the gap, it accelerates. It will gain a kinetic energy equal to the potential energy difference across the gap

$$\Delta K = q\Delta V$$

As the particle travels around the D to the other side of the, the cyclotron, the cyclotron switches the polarity of the potential difference. So as the particle passes the gap on the other side of the cyclotron, it is again accelerated with an additional $\Delta K = q\Delta V$. Since r does depend on the speed,

$$r = \frac{mv}{qB}$$

the radius increases with each “kick.” Finally the particle leaves the cyclotron with a velocity of

$$\frac{qBr_{\max}}{m} = v$$

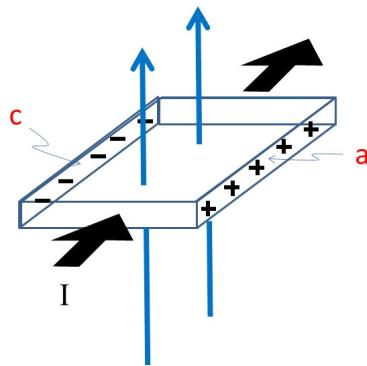
Since we often describe the velocity of particles in energy terms, the kinetic energy of the particle

$$\begin{aligned} K &= \frac{1}{2}mv^2 \\ &= \frac{1}{2}m\left(\frac{qBr_{\max}}{m}\right)^2 \\ &= \frac{q^2B^2r_{\max}^2}{2m} \end{aligned}$$

Hall Effect

Hall Effect Demo

The Hall probe is a cool little device that measures the magnitude of the magnetic field. It is used in rotation and angle detection in engineering. We should find out how it works.



Let's take a piece of material that has a current going through it. If we place it in a magnetic field, then the charge carriers will feel a force. Suppose it is a metal, and that the charge carriers are electrons. The force is perpendicular to the current direction. So the electrons are accelerated toward the top of the piece of metal as shown in the drawing. This creates a negative charge on the top side of the metal piece. Then the bottom side will be positively charged relative to the top. With separated charge like this, we think of a capacitor and the electric field created by such a separation of charges. There will be a field in the conductor with a potential difference between the

top and bottom of the conductor. We call this potential difference

$$\Delta V_H$$

the Hall potential after the man who first observed it.

Now if the charge carriers were positive, we would still build up a potential, but it would be in the opposite polarity. We wish to find this hall potential. The electric field of the charges will try to push them back down as more charge builds up. So at some point the upward force due to the magnetic field on the electrons will be balanced by the built up electric field. At that point

$$\Sigma F_y = 0 = F_B - F_E$$

so

$$qv_d B = qE_H$$

where E_H is the field due to the separation of charges.

So

$$E_H = v_d B$$

The potential is nearly equal to

$$\Delta V \approx E_H d$$

where d is the top-to-bottom distance of the conductor , so

$$\Delta V \approx v_d B d$$

Since we know

$$I = nqAv_d$$

then

$$v_d = \frac{I}{nqA}$$

and the area A is

$$A = td$$

where t is the thickness of the conductor, then

$$v_d = \frac{I}{nqtd}$$

and

$$\Delta V \approx \frac{IB}{nqt}$$

You may find this expressed in terms of the Hall coefficient

$$R_H = \frac{1}{nq}$$

so

$$\Delta V \approx R_H \frac{IB}{t}$$

To do a good job of finding R_H for metals and semiconductors, you have to go beyond classical theory. But if we know B , I , t , and ΔV , which can all be measured, then we

can find R_H . Once this is done, we can place the Hall probe in different magnetic fields to find their strength. One way to do this is to control I and measure ΔV , so

$$B \approx \frac{t}{R_H I} \Delta V$$

Basic Equations

41 Magnetic forces on wires

Fundamental Concepts

- The magnetic force on moving charges extends to wires with currents
- The force on a wire with current is given by $\mathbf{F}_I = I\mathbf{L} \times \mathbf{B}$
- The torque on a current loop is $\tau = \mu \times \mathbf{B}$ where $\mu = IA$

Magnetic forces on Current-Carrying wires

Question 223.41.1

Question 223.41.2

If there is a force on a single moving charge due to a magnetic field, then there must be a force on lots of moving charges! We call lots of moving charges an electric current

$$I = \frac{\Delta Q}{\Delta t}$$

For charges in a wire, we know that the charges move along the wire with a velocity v_d . We would expect the total force on all the charges to be the sum of all the forces on the individual charges.

$$F_I = \sum_i F_{q_i} = \sum_i q_i v B \sin \theta$$

but, since in our wire all the charge carriers are the same, this is just

$$F_I = N q_i v_d B \sin \theta$$

where here N is the number of charge carriers in the part of the wire that is experiencing the field. We used a charge density n before. Let's use it again to make an expression for N

$$N = nV = nAL$$

where A is the cross sectional area of the wire and L is the length of the wire. So

$$F_I = nALq_i v_d B \sin \theta$$

Now let's think back to our definition of current. We know that

$$I = nq_i v_d A$$

so our force on the current carrying wire is

$$\begin{aligned} F_I &= (nqv_d A) LB \sin \theta \\ &= ILB \sin \theta \end{aligned}$$

Remember that θ is the angle between the field direction and the velocity. In this case I is in the direction of the velocity (we still assume positive charge carriers, even though we know they are electrons going the other way). So θ is the angle between the field direction and the direction of the current. We can write this as a cross product

$$\vec{F}_I = I \vec{L} \times \vec{B} \quad (41.1)$$

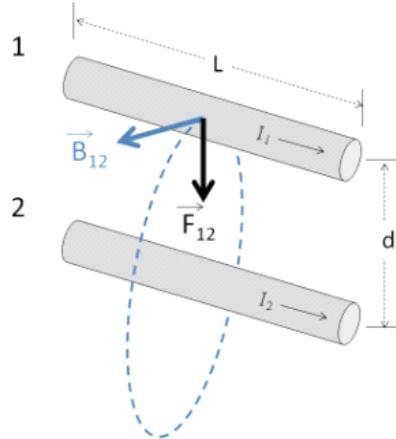
where \vec{L} is in the current direction.

Force between two wires

Question
223.41.2.3

We can use what we have learned to find the force between two wires.

If I have two wires with current, I will have a field created by each wire. Let's suppose that I_1 and I_2 are in the same direction



and let's calculate the force on wire 1 due to the field of wire 2. The field due to wire 2 at the location of wire 1 will be

$$B_{12} = \frac{\mu_0 I_2}{2\pi d}$$

where d is how far away wire 1 is from wire 2. We know

$$F_{12} = I_1 L B_{12} \sin \theta$$

We can see that $\sin \theta = 1$ since I_1 will be perpendicular to B_{12} .

$$F_{12} = I_1 L B_{12}$$

and using our expression for B_{12}

$$\begin{aligned} F_{12} &= I_1 L \frac{\mu_o I_2}{2\pi d} \\ &= L \frac{\mu_o I_2 I_1}{2\pi d} \end{aligned} \quad (41.2)$$

Would you expect F_{21} to be very different?

Torque on a Current Loop

Question 223.41.4

Question 223.41.5

Remember that in PH121 or Statics and Dynamics we defined angular displacement

$$\Delta\theta = \theta_f - \theta_i \quad (41.3)$$

and this told us how far in angle we had traveled from a starting point θ_i .

We also defined the angular velocity

$$\omega = \frac{\Delta\theta}{\Delta t} \quad (41.4)$$

which told us how fast an object was spinning in radians per second. The direction of this angular velocity we found using a right hand rule.

We also defined an angular acceleration

$$\alpha = \frac{\Delta\omega}{\Delta t} \quad (41.5)$$

and we used angular acceleration in combination with a moment of inertia to express a rotational form of Newton's second law

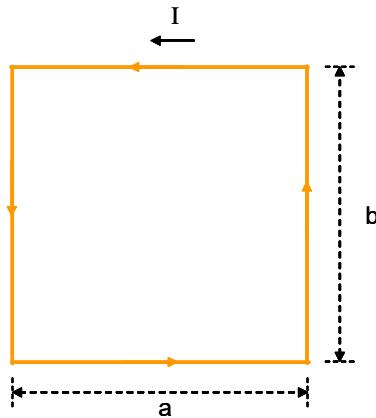
$$\sum \tau = I\alpha \quad (41.6)$$

where τ is a torque. We found torque with the expression

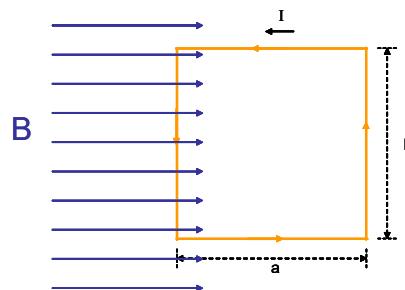
$$\vec{\tau} = \vec{r} \times \vec{F} \quad (41.7)$$

We wish to apply these ideas to our new force on wires due to magnetism.

Let's take a specific example. I want to use a current loop. This is just the simple loop of current we have seen before.



I want to place this into a magnetic field.



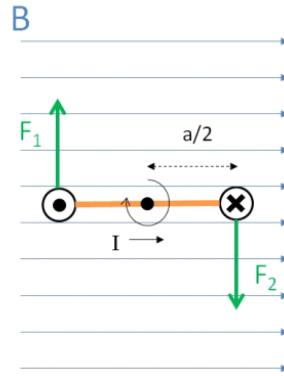
I drew the current loop as a rectangle on purpose, I want to look at the force on the current for each part of the loop. Each side of our loop is a straight wire segment. Remember that the magnitude of the force on a wire is given by

$$F_I = ILB \sin \theta$$

where θ is the angle between I and B so if $\theta = 0$ or if $\theta = \pi$ rad, then $\sin \theta$ will be zero. The magnitude of the force will then be zero. So the top and bottom parts of the loop will not experience a force. The sides will, though, and since for $\theta = \frac{\pi}{2}$ or $\theta = -\frac{\pi}{2}$ ($\theta = -\frac{\pi}{2}$ is the same as $\theta = \frac{3\pi}{2}$) then $\sin \theta = 1$ and the force will be a maximum.

$$F_I = IbB$$

on each side wire segment. But we need to consider direction. The force will be perpendicular to both I and B . We use our right hand rule. Fingers in the direction of I , curl to the direction of B . We see the force is out of the figure for the left hand side and into the figure for the right hand side. The next figure is a bottom-up view.



Clearly the loop will want to turn! This looks like a nice problem for us to describe with a torque. We have a force acting at a distance from a pivot. We have a torque

$$\tau = rF \sin \psi$$

We have already used θ , and our torque angle is the angle between r and F , so we needed a new greek letter. I have used ψ ³². Then ψ is the angle between r and F .

Let's fill in the details of our total torque. Remember we have two torques, one for the left hand side, and one for the right hand side. Their magnitudes are the same, and the directions we need to get from yet another right hand rule. Both are in the same direction so

$$\begin{aligned}\tau &= \frac{a}{2} F_I \sin(\psi) + \frac{a}{2} (F_I) \sin(\psi) \\ &= aF_I \sin(\psi)\end{aligned}$$

Putting in the force magnitude gives

$$\tau = a(IbB) \sin \psi$$

and rearranging lets us see

$$\begin{aligned}\tau &= (ab) IB \sin \psi \\ &= (A) IB \sin \psi\end{aligned}$$

where $A = ab$ is the area of our loop. Of course we can write this as

$$\vec{\tau} = I \vec{A} \times \vec{B} \quad (41.8)$$

The torque is the cross product of the area vector and the magnetic field multiplied by the current.

We did this for a square loop. It turns out that it works for any loop shape.

When things rotate, we expect to use moments. We defined a magnetic dipole moment

³² which is a *psi*

for a current loop. Now we can see why it is useful. The magnetic moment tells us about how much torque we will get for a particular current loop.

$$\vec{\mu}_d = I \vec{A}$$

using this we have

$$\vec{\tau} = \vec{\mu}_d \times \vec{B}$$

We could envision our loop as a single circle of wire connected to a battery. But we could just as easily double up the wire. If we do this, what is our torque? Well we would have twice the force, because we now have twice the current (the current goes through both turns of the wire). So now we have

$$\tau = 2(A)IB \sin \psi$$

But why stop there? We could make three loops all together.

$$\tau = 3(A)IB \sin \psi$$

or many more, say N loops,

$$\tau = NAIB \sin \psi$$

Thinking of our magnetic dipole moment, we see that

$$\tau = N\mu_d B \sin \psi$$

for a coil. We could combine the effects of all the loops into one magnetic moment that represents the coil.

$$\vec{\mu} = N \vec{A} I \quad (41.9)$$

then

$$\tau = \mu B \sin \psi$$

or in cross product form

$$\vec{\tau} = \vec{\mu} \times \vec{B} \quad (41.10)$$

Using this total magnetic moment, we can more easily do problems with coils in magnetic fields.

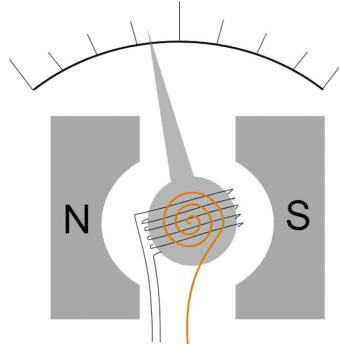
For example, we found that there was a potential energy associated with spinning dipoles, for a spinning current loop we also expect a potential energy. We have a simple formula for this potential energy in terms of the magnetic moment.

Question 223.41.5

$$U = -\vec{\mu} \cdot \vec{B} \quad (41.11)$$

Galvanometer

We finally know enough to understand how to measure a current. The device is called a *galvanometer*.

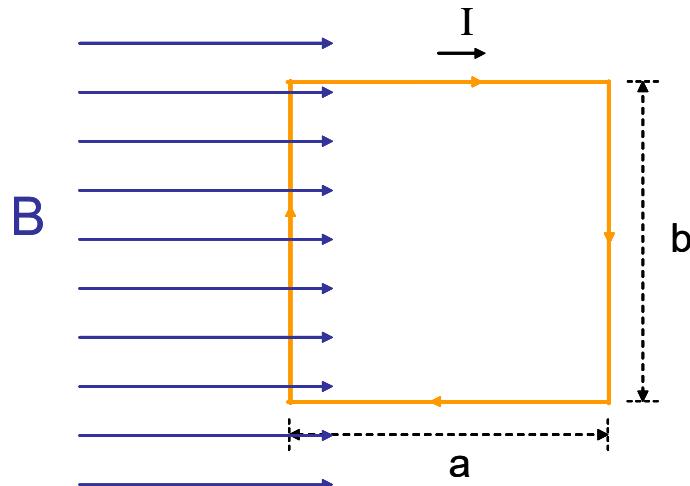


In the picture, we see the typical design of a galvanometer. It has a coil of wire (shown looking at the side of the coil) and a spring. The coil is placed between the ends of a magnet. When there is a current in the wire, there will be a torque on the coil that will compress the spring. The amount of torque depends on the current. As the current increases, the spring is more compressed. A marker (large needle) is attached to the apparatus. As the spring is compressed, the indicator moves across the scale. Since this movement is proportional to the current, a galvanometer can easily measure current.

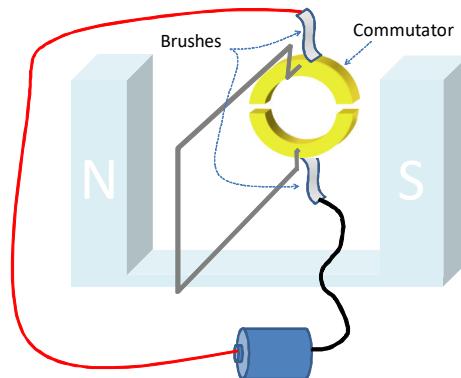
Electric Motors

Question 223.41.6

With our new understanding of torque on a current loop, we should be able to see how an electric motor works. A current loop is placed in between two magnets to form a magnetic field. The loop will turn because of the torque due to the B -field. But we have to get clever. What happens when the loop turns half way around so the current is now going the opposite way?



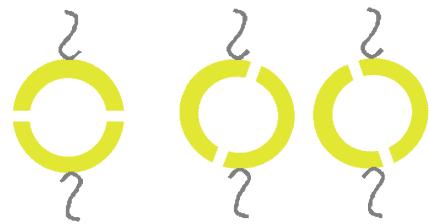
Now the torque switches direction and the loop will come to rest. We don't want that if we are building a motor, so we have to switch the current direction every time the loop turns half way.



The way we do this is to have electrical contacts that are flexible, called brushes. The brushes contact a metal ring. The metal ring is connected to the loop. But the ring has two slits cut out of it.



The ring with slits is called a commutator. As the loop turns, the commutator turns, and when it has turned a half turn, the brushes switch sides. This changes the current direction, which puts us back at maximum torque.



This keeps the motor going the same direction.

Basic Equations

42 Permanent Magnets, Induction

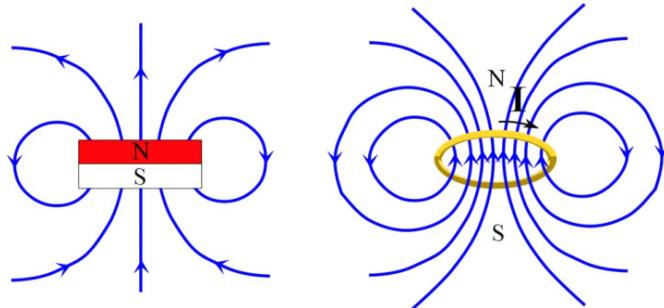
Fundamental Concepts

- Using classical physics, we can't quite explain a permanent magnet.
- Using a semiclassical model, the permanent magnet's field is due to spinning electrons.
- Alignment of the spinning electrons creates what we call magnetism.
- Temporary alignment results in paramagnetism and diamagnetism.
- More permanent alignment yields ferromagnetism.
- A changing magnetic field can create an emf.

Finally, why magnets work

We started our study of magnetism by looking at bar magnets and considering that if we break one, we end up with two magnets. We don't end up with a north end and a south end as separate pieces. This is different than charge and electric fields. Then we studied how moving charge makes a magnetic field. But we didn't say how bar magnets work yet. Can a permanent magnet have something to do with current loops?

Question 223.42.1



Well, lets look at the field due to a current loop. It looks a lot like the field due to a magnet. Could there be current loops inside a bar magnet? The answer is well, sort

of... We have electrons that sort of travel around the atom. Suppose the electrons orbit like planets. Then there would be a current as they travel. For one electron the current would be

$$I = \frac{q_e}{T}$$

where T is the period of rotation. And we recall from PH121 or Statics and Dynamics that the period of rotation can be found from

$$\omega = \frac{2\pi}{T}$$

so that

$$T = \frac{2\pi}{\omega}$$

Then the current is

$$I = \frac{q_e \omega}{2\pi}$$

It is an amount of charge per unit time. We can write this as

$$I = q_e \frac{\omega}{2\pi}$$

and recalling

$$v_t = \omega r$$

then

$$I = q_e \frac{v_t}{2\pi r}$$

We can find a magnetic moment (a good review of what we have learned!)

$$\begin{aligned} \mu &= NIA = (1) IA \\ &= q_e \frac{v_t}{2\pi r} (\pi r^2) \\ &= \frac{q_e v_t r}{2} \end{aligned}$$

Physicists often write this in terms of angular momentum. Just to review, angular momentum is given by

$$L = I_m \omega$$

where I_m is the moment of inertia. Then

$$\begin{aligned} L &= I_m \omega \\ &= (mr^2) \left(\frac{v_t}{r} \right) \\ &= mr v_t \end{aligned}$$

so the magnetic moment of the orbiting electron would be

$$\mu = \frac{q_e L}{2m} \quad (42.1)$$

which gives us a magnetic moment related to the angular momentum of the electron. And if we have a magnetic moment this not only means the atoms would orient in an external field but it also means that the atoms work as little magnets. We will have a magnetic field

Quantum effects

Question 223.42.2

All of this works well for Hydrogen. We find that individual hydrogen atoms do act like small magnets. But if the hydrogen is in a compound, it is more complicated because we then have many electrons and they are “orbiting” in different directions. It is even true that most atoms have many electrons, and within the atom these electrons fly around in all different directions. The magnetic field due to one electron in the atom cancels out the magnetic field due to another, so there is no net magnetic field due to the “motion” of the electrons in their orbitals. So in general there is no net magnetic field from orbital motion. Even for Hydrogen in a compound the overall magnetic moment of the compound tends to cancel out.

Further, we know that electrons do not travel like planets in circular orbits. So our model for magnetism is not really correct yet.³³ To understand the current model of electron orbitals takes some quantum mechanics (and a few more years of physics). But we can understand a little, because quantum mechanics does tell us that the electrons have angular momentum. The big difference is that the angular momentum is *quantized* meaning it can have only certain values (think of the quantized modes of an oscillating string). The smallest magnetic moment for an electron turns out to be

$$\mu = \sqrt{2} \frac{q_e}{2m_e} \hbar \quad (42.2)$$

where

$$\hbar = \frac{h}{2\pi} = 1.05 \times 10^{-34} \text{ Js}$$

is pronounced “h-bar” and is a constant. We encountered Planck’s constant h before ($h = 6.63 \times 10^{-34} \text{ Js}$). This is just Planck’s constant divided by 2π . So it would seem that with only certain values being available the magnetic moments might be more likely to line up.

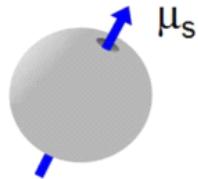
But it turns out that even in quantum mechanics, the magnetic moments of the electrons due to their orbits cancel each other out most of the time.

But there is another contribution to the magnetic moment, this time from the electron, itself. The electron has an amount of angular momentum. It is as though it spins on an axis. This spin angular momentum is also quantized. It can take values of

$$S = \pm \frac{\sqrt{3}}{2} \hbar \quad (42.3)$$

³³ Strictly speaking, the electrons don’t move like little planets around the nucleus. So it is not clear that orbital “motion” makes much sense. The electrons form standing waves around the nucleus that oscillate in time. But they do have angular momentum, and that orbital angular momentum cancels out for most atoms. For more on this, take PH279, Modern Physics.

My mental picture of this is a charged ball spinning on an axis.³⁴



The magnetic moment due to spin is

$$\mu_s = \frac{q_e \hbar}{2m_e} \quad (42.4)$$

This means that electrons, themselves are little magnets. Where does this magnetic moment come from? Well it is *as though* the electron is constantly spinning. It is not really, but this is a semi-classical mental model that we can use to envision the source of the electron's magnetic field. The “spinning” electron is charged, so the electron acts like a minuscule current loop. The electron, itself is a source of the magnetic field for permanent magnets.

The spin magnetic moment was given the strange name *Bohr magneton* in honor of Niels Bohr. If there are many electrons in the atom, there will be many contributions to the total atomic magnetic moment. The nucleus also has a magnetic moment (a detail we will not discuss at any length in our class) and there are other details like electron spin states pairing up. But those are topics for PH279 and our senior quantum mechanics class. It turns out that this spin magnetic moment is the major cause that produces permanent magnetism in some metals. We don't want to wade though a senior level physics class now (well, you probably don't anyway) so we need a more macroscopic description of magnetism. But fundamentally, if we can get the electrons spins in a material to line up, we will have a magnet.

Ferromagnetism

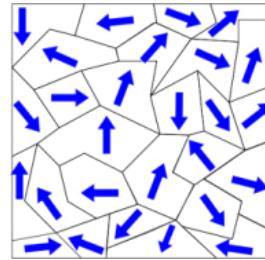
Question 223.42.3

Because of the spin magnetic moment, we can see some hope for how a permanent might work. But these spin magnetic moments are also mostly randomly arranged. So again, most atoms won't have an overall magnetic moment. But some atoms do have a slight net field. They have an odd number of electrons. So the last electron can have an unbalanced magnetic moment. That atom would act as a magnet

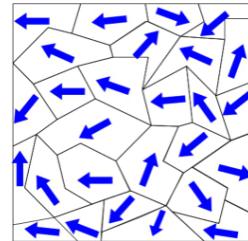
Still, this does not produce much of an effect, because neighboring atoms all are

³⁴ But this is just a mental model. Electrons don't spin. They do have angular momentum. But how electrons generate their angular momentum is still hard to tell. Physicists are working on this.

oriented differently. So neighboring atoms cancel each other out. In a few materials, though, the atoms within small volumes will align their magnetic moments. These little domains form small magnets. But still the overall effect is very small because the domains are all oriented in different directions.



If we place these materials in a magnetic field, we can make the domains align, and then we have something!



Few materials can do this. The ones that can are called ferromagnetic. Iron is one material. We can make the domains align, but the alignment decays quickly. That is why iron objects stick to a magnet, but don't stick to each other when they are taken away from the magnet. But if we can force the domains to stay in one direction, say, by heating the ferromagnetic metal in a magnetic field and letting it cool and form crystals, then we can make a magnet that will last longer. The magnetic moments will get stuck all pointing about the same direction as the ferromagnetic metal cools. Some materials like Cobalt form very long lasting permanent magnets.

Magnetization vector

We now know that each atom of a substance may have a magnetic moment. For a block of the material, it is useful to think of the magnetic moment per unit volume. We will call this \mathbf{M} . It must be a vector, so that if there is an overall magnetic moment, we have a magnet! Let's see how to use this new quantity.

Suppose I have a current carrying wire that produces a field \mathbf{B}_o . But I also have a material where \mathbf{M} is not zero. Then there must be a field due to the magnetic material

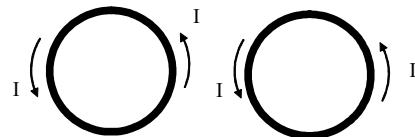
\mathbf{B}_m . So the total field will be

$$\mathbf{B} = \mathbf{B}_o + \mathbf{B}_m \quad (42.5)$$

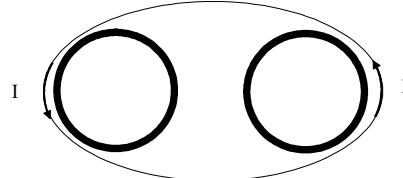
and all we have to do is determine the relationship between \mathbf{B}_m and \mathbf{M} .

Solenoid approximation

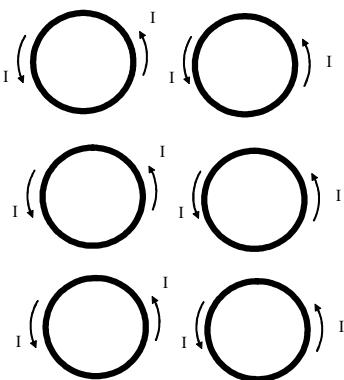
Lets look at two atoms, We will model them as little current loops, since they have magnetic moments.



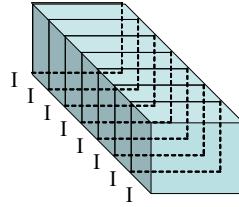
notice that in between the loops, the currents go opposite directions. We could think of them as canceling. We get a net current that is to the outside of the loops



Now let's take many current loops.



again, the inside currents cancel, leaving an overall current along the outside. Now if we view a material as a stack of such current loops



we can model a magnetic material like a solenoid! That is great, because we know how to find the field of a solenoid.

$$\begin{aligned} B_m &= \mu_o n I \\ &= \mu_o \frac{NIA}{\ell A} \end{aligned}$$

I didn't cancel the A s because I want to recognize the numerator as the magnetic moment

$$\mu = NIA$$

so

$$B_m = \mu_o \frac{\mu}{\ell A}$$

But note that ℓA is just the volume of the piece of magnetic material, so

$$B_m = \mu_o \frac{\mu}{V}$$

which gives us our new quantity, the magnetization vector

$$M = \frac{\mu}{V} \tag{42.6}$$

Well, this is the magnitude, anyway, so

$$B_m = \mu_o M \tag{42.7}$$

and of course the directions must be the same, since μ_o is just a scalar constant

$$\mathbf{B}_m = \mu_o \mathbf{M} \tag{42.8}$$

So the total field is given by

$$\mathbf{B} = \mathbf{B}_o + \mu_o \mathbf{M} \tag{42.9}$$

Magnetic Field Strength (another confusing name)

Only do this if you have extra time

Sometimes we physicists just can't let things alone. So when we arrived at the equation

$$\mathbf{B} = \mathbf{B}_o + \mu_o \mathbf{M} \tag{42.10}$$

someone wanted to define a new term

$$\frac{\mathbf{B}_o}{\mu_o} \quad (42.11)$$

so we could write the equation

$$\mathbf{B} = \mu_o \left(\frac{\mathbf{B}_o}{\mu_o} + \mathbf{M} \right) \quad (42.12)$$

This new term is given an unfortunate name. The *magnetic field strength*. **It is not the magnitude of the magnetic field**, but is the magnitude divided by the constant μ_o . It has its own symbol, \mathbf{H} . So you may write our total field equation as

$$\mathbf{B} = \mu_o (\mathbf{H} + \mathbf{M}) \quad (42.13)$$

You might find this change unnecessary and confusing (I do) but it is tradition to use this notation, and is not bad once you get used to it.

Macroscopic properties of magnetic materials

We want a way to describe how “magnetic” different substances are without doing quantum mechanics. This will allow us to classify materials, and choose the proper material for whatever experiment or device we are designing.

For many substances we find that the magnetization vector is proportional to the field strength (which is why field strength hangs around in usage)

$$\mathbf{M} = \chi \mathbf{H} \quad (42.14)$$

For many materials, this nice linear relationship applies, and we can look up the constant of proportionality in a table. The name of the constant χ is the *magnetic susceptibility*.

If χ is positive (M is in the same direction as H), we call the material *paramagnetic*.

If χ is negative (M is in the opposite direction as H), we call the material *diamagnetic*.

Using this new notation, our total field becomes

$$\begin{aligned} \mathbf{B} &= \mu_o (\mathbf{H} + \mathbf{M}) \\ \mathbf{B} &= \mu_o (\mathbf{H} + \chi \mathbf{H}) \end{aligned}$$

$$\mathbf{B} = \mu_o (1 + \chi) \mathbf{H} \quad (42.15)$$

The quantity $\mu_o (1 + \chi)$ is also given a name,

$$\mu_m = \mu_o (1 + \chi) \quad (42.16)$$

it is called the magnetic permeability. Now you see why μ_o is called the permeability of free space! (the name was not so random after all!). If $\chi = 0$ then

$$\mu_m = \mu_o \quad (42.17)$$

and this is the case for free space. We can write definitions of paramagnetism and diamagnetism in terms of the permeability.

Paramagnetic	$\mu_m > \mu_o$
Diamagnetic	$\mu_m < \mu_o$
Free Space	$\mu_m = \mu_o$

For paramagnetic and diamagnetic materials, μ_m is usually not too different from μ_o but for ferromagnetic materials μ_m is much larger than μ_o . Note that we have not included ferromagnetic substances in this discussion. That is because

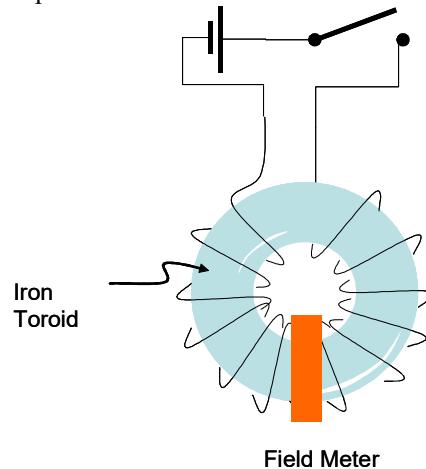
$$\mathbf{M} = \chi \mathbf{H}$$

is not true for ferromagnetic materials.

Ferromagnetism revisited

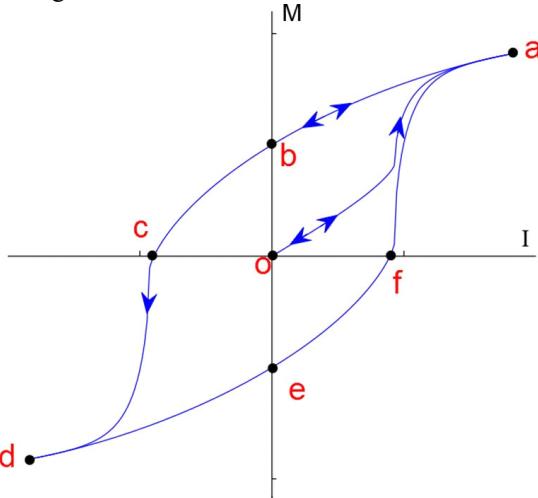
Question 223.42.4

But why is ferromagnetism different? To try to understand, let's take a iron toroid (doughnut shape) and wrap it with a coil as shown.



We have a magnetic field meter that measures the field inside the windings of the coil. When we throw the switch, the coil produces a magnetic field. The field will produce a magnetization vector in the iron toroid and, therefore, a field strength. We can plot the applied magnetic field vs. the field strength to see how much effect the applied field has on the magnetic properties of the iron toroid. We won't do this mathematically, but the

result is shown in the figure.



As we throw the switch, we go from no alignment of the domains so zero M and therefore zero induced field in the iron toroid to a value that represents the almost complete alignment of the magnetic moments of each atom of the iron. This is point *a*. It may take a bit of current, but in theory we can always do this. All the domains are aligned and M is maximum.

Now we reduce the current from our battery, and we find that the field due to the aligned domains drops as expected, but not along the same path that we started on! We go from *a* to *b* as the current decreases. At point *b* there is no current, but we still have a magnetic field in the toroid!

We can even keep going and reverse the field by changing the polarity of our power supply contacts. Since we still have some field in the toroid, it actually takes some reverse current to overcome the residual field. But if we apply enough reverse current, then we get alignment in the other direction. Almost complete alignment is at point *d*. If we again reduce the current and find that—once again—it does not retrace the same path!

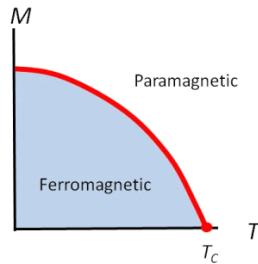
Each time we align the domains with our applied external field from the coil, the domains in the iron toroid seem to want to stay aligned. Most do lose alignment, but some stay put. We have created a weak permanent magnet by placing our ferromagnetic material in a strong external magnetic field.

This strangely shaped curve is the *magnetization curve* for the material. The fact that the path is a strange loop instead of always following the same path is called *magnetic hysteresis*. We can see now that the external field (represented by the current I , since

$B_{external} \propto I$) and magnetization don't behave in a simple relationship like they did for diamagnetic or permanganic materials.

The thickness of the area traced by the hysteresis curve depends on the material. It also represents the energy required to take the material through the hysteresis cycle.

If we add enough thermal energy, it is hard to keep the atomic dipole moments aligned. The next figure shows this effect.



At a temperature called the Curie temperature, the material no longer acts ferromagnetic. It becomes simply paramagnetic. So if we heat up a permanent magnet, we expect it to lose its alignment and therefore to stop being a magnet. This is what happens to ferromagnetic materials when they are heated due to volcanism. The domains are destroyed and all the atoms lose alignment. When the material cools, the Earth's magnetic field acts as an external field and some of the domains will be aligned with this field. This is how we know that the Earth's magnetic field switches polarity. We can see which way the magnetization vector points in the cooled lava deposits from places like the Mid-Atlantic Trench.

This is also how old fashioned magnetic tapes and disks work.

Paramagnetism

We said that if $\mu_m > \mu_o$ we get paramagnetism. But what is paramagnetism? It comes from the material having a small natural magnetic susceptibility.

$$0 < \chi \ll 1 \quad (42.18)$$

So in the presence of an external magnetic field, you can force the little magnetic moments to line up. You are competing with thermal motion as we saw in ferromagnetism, so the effect is usually weak. A rule of thumb for paramagnetism is

that

$$M = C \frac{B_o}{T} \quad (42.19)$$

where C contains the particular material properties of the substance you are investigating (another thing to look up in tables in data sets), B_o is the applied field, and T is the temperature. In other words, if it is cool enough, a paramagnetic material becomes a magnet in the presence of an external magnetic field. This is a little like polarization of neutral insulators in the presence of an electric field. For paramagnetic materials, the induced magnetic field is in the same direction as the external field.

Some examples of paramagnetic materials and their susceptibilities are given below

Material	Susceptibility
Tungsten	6.8×10^{-5}
Aluminium	2.2×10^{-5}
Sodium	0.72×10^{-5}

Diamagnetism.

If $\mu_m < \mu_o$ we said we would have diamagnetism. This is fundamentally quite different from paramagnetism. It comes from the material having paired electrons that orbit the atom (classical model). The magnetic moments of the electrons will have equal magnitudes, but opposite directions (a little bit of quantum mechanics to go with our classical model). When the external field is applied, one electron's orbit is enhanced by the field, and the other is diminished (think $q\mathbf{v} \times \mathbf{B}$). So there will be a net magnetic moment. If you think about this for a while, you will realize that the new net magnetic moment is in the opposite direction of the applied external field! So diamagnetism will always repel.

There is always some diamagnetism in all matter. We can enhance the effect using a superconductor. The diamagnetism of the superconductor repels the external field entirely! Why does this happen only for superconductors? Well, that will take more theory to discover (a great topic for our junior level electrodynamics class). But the

Meissner effect
demo

Back to the Earth

So now we can see that the Earth is a magnet and we know how magnets are formed. But wait, why is the Earth a magnet? The real answer is that we don't know. But we believe that again it is because of current loops. We believe there is a current of ionized

Nickel and Iron in near the center of the Earth. So the flow of these charged liquid metals will create a magnetic field. This is a very large current loop! The evidence for this is that magnetic field seems proportional to the spin rate of the planet. But this is an area of active research.

It is curious that the magnetic pole and the geographic pole are not in the same place. The magnetic pole also moves around like a precession. Then, every couple of hundred thousand years, the polarity of the Earth's field switches altogether!

There is still plenty of good research to do in this area.

The location of the magnetic pole explains the declination adjustment you have to use when using a compass. What you are really doing is accounting for the difference in pole location.

Induced currents

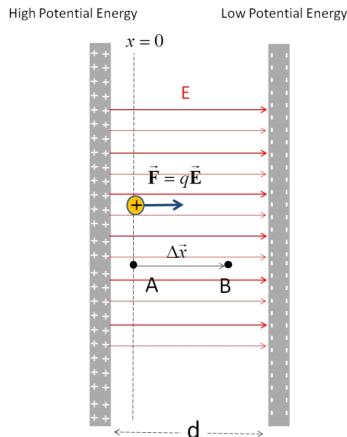
Question 223.42.5

We spend most of the last two lectures building a relationship between moving charge (current) and magnetic fields. But suppose we have moving magnetic fields. Could a moving magnetic field make a current?

If we think of relative motion, it seems like it should. After all, how do we know that it is the charge that is moving and not a moving B -field. In fact, moving B -fields *do* cause a current. We say that a moving or changing magnetic field *induces* a current.

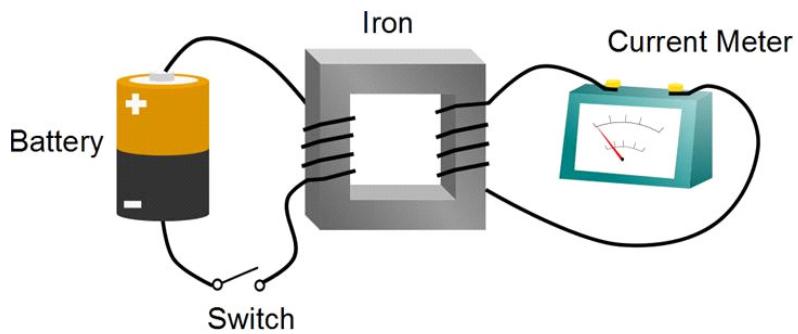
Faraday discovered this effect. He described it as an *induced emf*. An emf is something that "pumps" the charges in the wire. It takes them from a lower to a higher potential so they can form a current. The changing magnetic field must be "pumping" the charges as it changes!

What is really going on here? Think for a minute what must be happening.



When we defined the electric potential, we use a capacitor. We found that there was a field directed from the + charges to the - charges. And in this field, charges had an amount of potential energy. When a current flows from the + end of the battery to the - end, there must be an electric field acting on the charge in the wire! That is what creates the electric potential. So, then, does a moving magnetic field create an electric field?

The answer is yes! We say that an electric field is *induced* by a moving magnetic field. This is really the same as saying that there is an induced emf for our current loop.



Faraday actually set up his experiment with two coils of wire. One coil was connected to a battery. We now know this coil will make a magnetic field. As the current starts flowing the field will form. While it is forming, it will induce an emf in the second coil. But this is just using an electromagnet instead of a permanent magnet.

To be able to calculate how much current flows, we will need to investigate changing magnetic fields. We will do this next lecture with our concept of flux.

Basic Equations

43 Induction

Fundamental Concepts

- Conductors moving in magnetic fields separate charge, creating a potential difference that we call “motional emf.”
- Motional emfs generate currents, even in solid pieces of conductor. These currents in conductors are called “eddy currents.”
- Magnetic flux is found by integrating the dot product of the magnetic field and a differential element of area over the area. $\Phi_B = \int_A \vec{B} \cdot d\vec{A}$

Motional emf

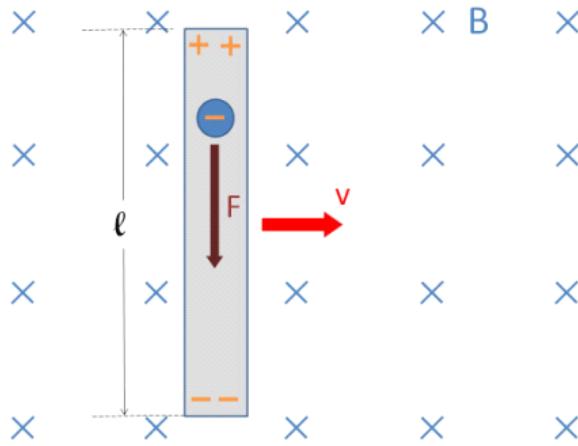
Last lecture, we studied Faraday’s experiment. He created a magnetic field, and then used that magnetic field to make a current. But currents are caused by electric fields! Did Faraday’s magnetic field create an electric field?

To investigate Faraday’s result, let’s see if we can find a way to use charge motion and a magnetic field to make an electric field. Let’s take a bar of metal and move it in a magnetic field. The bar has free charges in it (electrons). We have given them a velocity. So we expect a magnetic force

$$\vec{F}_B = q\vec{v} \times \vec{B}$$

The free charges will accelerate together, but the positive stationary charges can’t move. We have found another way to separate charge. We know that separated charge creates a potential difference. We often call this induced potential difference the *motional emf* because it is created by moving our apparatus.

Let’s take an example to see how it works.



For this example, let's look at a piece of wire moving in a constant field. To make the math easy, let's move the wire with a velocity perpendicular to the B -field.

As the figure shows, the electrons will feel a force. Using our right hand rule, we get an upward force for positive charge carriers, but we know the electrons are negative charge carriers, so the force is downward. We find that the magnitude of the force is

$$F_B = qvB$$

The electrons will bunch up at the bottom of the piece of wire, until their electric force of repulsion forces them to stop. That force is

$$F_E = qE$$

By separating the charges along the wire so that there is excess positive charge on one end and excess negative charge on the other end, we now have an E -field in the wire. We can solve for E when we have reached equilibrium.

$$\Sigma F = 0 = -F_B + F_E$$

or

$$qE = qvB$$

which tells us

$$E = vB \quad (4.3.1)$$

Now, we know that electric fields cause potential differences. The E -field in the wire will be nearly uniform. Then it looks much like a capacitor with separated charges. The potential difference will be

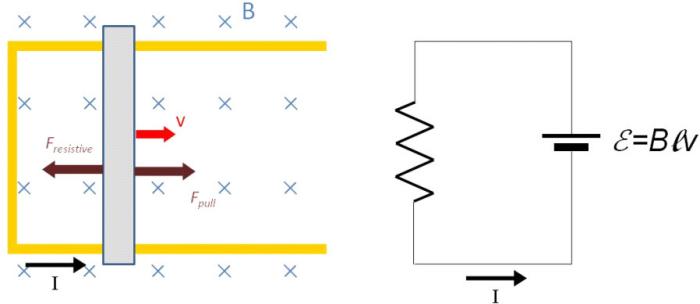
$$\begin{aligned} \Delta V &= \int \vec{E} \cdot d\vec{s} \\ &\approx EL \end{aligned}$$

where L is the length of our wire. So

$$\Delta V \approx vBL \quad (43.2)$$

This is like a battery. The magnetic field is “pumping” charge. If we connected the two ends somehow with a wire that is not moving, a current will flow (that is tricky to actually do!).

Question
223.43.0.1



Let's take another example. We wish to make a bar of metal move in a B -field. To make the rest of the circuit, we allow the bar to slide along two wires as shown. We will call the two wires “rails” since they look a little like railroad rails. Then we have a connection between our moving piece of metal, and the rest of the circuit. What we have is very like the circuit on the right hand side of the last figure.

We will have to apply a force F_{pull} to move the bar. This is because there is another force, marked as $F_{resistive}$ in the figure. This force is one we know, but might not recognize unless we think about it. We now have a current flowing through a wire, and the wire is in a magnetic field. So there will be a force

$$\begin{aligned} F_{resistive} &= I\vec{L} \times \vec{B} \\ &= ILB \sin \theta \\ &= ILB \end{aligned}$$

pushing to the left. This force resists our pull.

From Ohm's law, the current in the wire will be

$$\begin{aligned} I &= \frac{\Delta V}{R} \\ &= \frac{vBL}{R} \end{aligned}$$

so the force is

$$\begin{aligned} F_{resistive} &= \left(\frac{vBL}{R} \right) LB \\ &= \frac{vB^2L^2}{R} \end{aligned}$$

Thus we have to push with an equal force

$$F_{push} = F_{resistive}$$

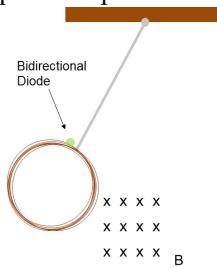
to keep the bar moving along the rails. If $F_{push} < F_{resistive}$ then the bar will have an acceleration, and it will be in the opposite direction from the velocity, so the bar will slow down.

Eddy Currents

Question 223.43.1

Pendulum-loop

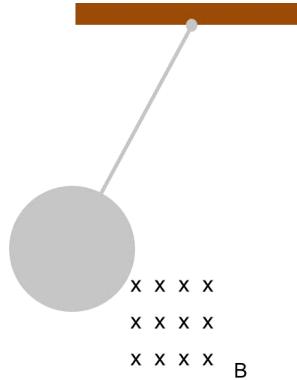
So if we have a conductive loop and part of that loop moves in a magnetic field, we get a current. I chose to make our apparatus a pendulum.



Pendulum-plate

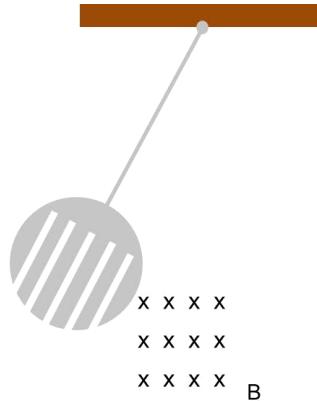
Question 223.43.2

So as the pendulum swings, through the magnetic field, we get a current. What if we have a solid sheet of conductor and we move that sheet through the magnetic field, will there be a current?



Question 223.43.3

The answer is yes. We call this current an *eddy current*. Let's see that this must be true with another experiment. Let's cut grooves in the plate.



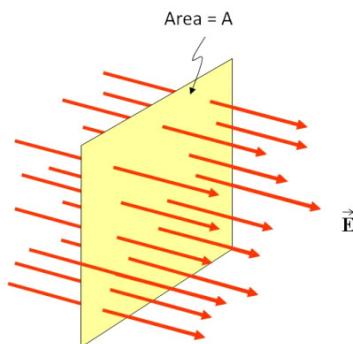
Al plate and strong magnets

Floating Plate Demo

The current is broken by the grooves, so there is little opposing magnetic field. This effect due to the eddy currents is often used to slow down machines. Rotating blades, and even trains use this effect to provide breaking.

Magnetic flux

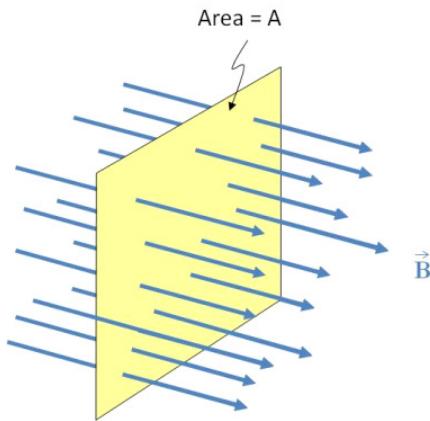
Remember long ago we defined the electric flux.



Recall that the electric flux is given by

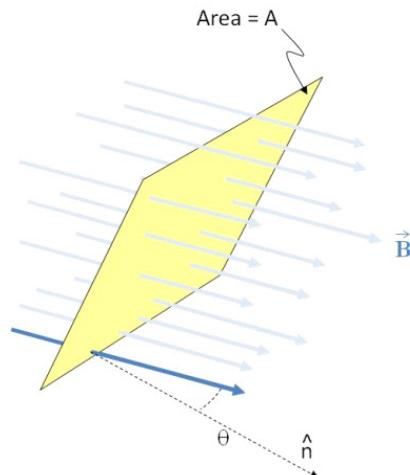
$$\begin{aligned}\Phi_E &= \vec{E} \cdot \vec{A} \\ &= EA \cos \theta\end{aligned}$$

But we now have a magnetic field.



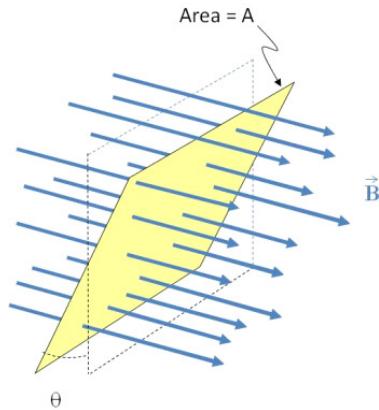
We define a magnetic flux

$$\Phi_B = \vec{B} \cdot \vec{A} \quad (43.3)$$



$$\Phi_B = BA \cos \theta \quad (43.4)$$

where θ is the angle between \vec{B} and \vec{A} .

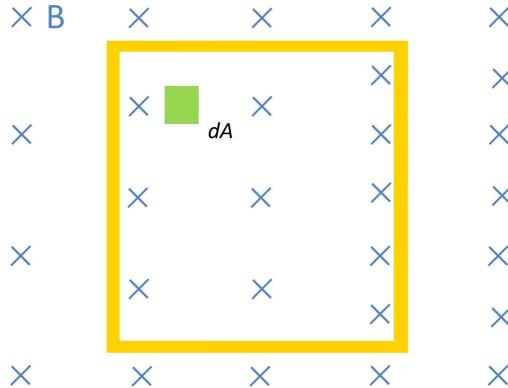


We found that the electric flux was very useful. We used Gauss' law to find fields using the idea of electric flux. It turns out that this magnetic flux is also a very useful idea. There is a difference, though. With electric fluxes, we had imaginary areas that the field penetrated. Often when we measure magnetic flux, we actually have something at the location of our area. We generally want to know the flux through a wire loop.

Just like with electric flux, we expect the flux to be proportional to the number of field lines that pass through the area.

Non uniform magnetic fields

So far in this lecture we have only drawn uniform magnetic fields and considered their flux. But we can easily imagine a non-uniform field. We tackled non-uniform electric field fluxes. We should take on non-uniform magnetic field fluxes as well. Suppose we have the situation shown in the following figure.



We have a loop of wire, and the loop is in a flux that changes from left to right.

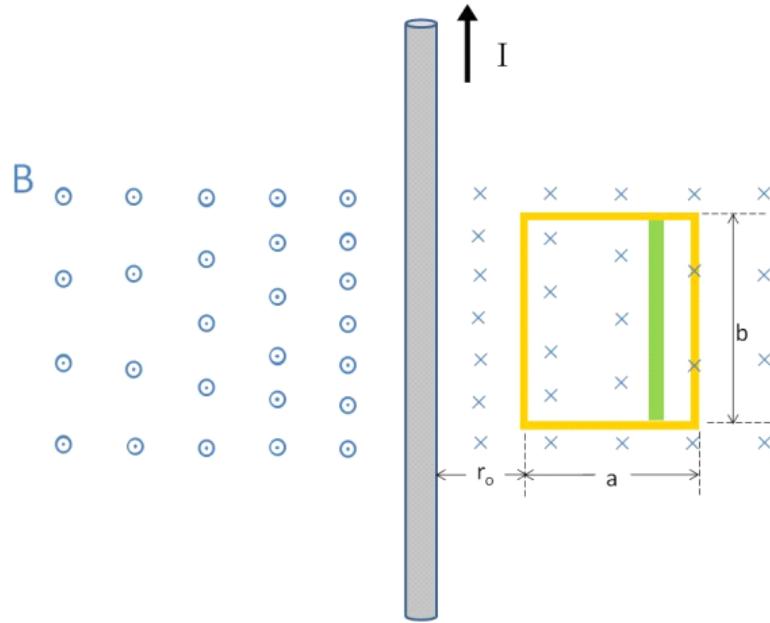
To find the flux through such a loop of wire, we can envision a small element of area, $d\vec{A}$ as shown. The flux through this area element is

$$d\Phi_B = \vec{B} \cdot d\vec{A}$$

We can integrate this to find the total flux

$$\Phi_B = \int_A \vec{B} \cdot d\vec{A} \quad (43.5)$$

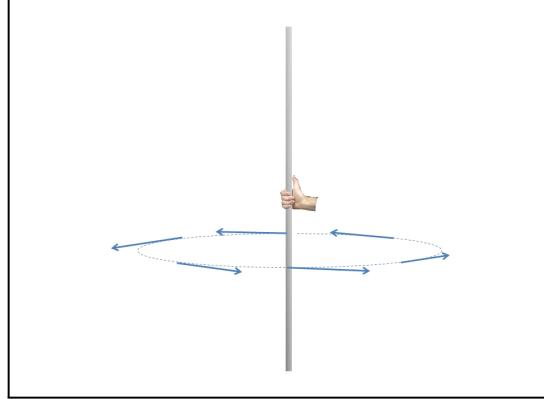
But what could make such a varying B -field? Consider a long straight wire again.



We know that the field due to the current-carrying wire will be

$$B = \frac{\mu_0 I}{2\pi r}$$

where r is the distance from the wire and the direction is given by one of our right hand rules.



Question 223.43.4

Question 223.43.5

The flux through the green rectangular area is almost constant. The little area is given by

$$dA = dydr$$

The area is perpendicular to the field, so the angle between B and A is 90° . Then

$$d\Phi_B = \frac{\mu_o I}{2\pi r} dy dr \quad (1)$$

and we can integrate this to find the total flux

$$\begin{aligned} \Phi_B &= \int_{r_o}^{r=a} \int_o^b \frac{\mu_o I}{2\pi r} dy dr \\ &= \frac{\mu_o I}{2\pi} \int_{r_o}^{r=a} \frac{1}{r} dr \int_o^b dy \\ &= \frac{\mu_o Ib}{2\pi} \int_{r_o}^{r_o+a} \frac{1}{r} dr \\ &= \frac{\mu_o Ib}{2\pi} (\ln(r_o + a) - \ln(r_o)) \\ &= \frac{\mu_o Ib}{2\pi} \ln\left(\frac{r_o + a}{r_o}\right) \end{aligned}$$

We can even put in some numbers for this case. Suppose our loop has a height of $b = 0.05$ m and a width of $a = 0.01$ m and that it is a distance $r_o = a$ away from the current carrying wire and that the current is $I = 0.5$ A. Then

$$\begin{aligned} \Phi_B &= \frac{(4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}})(0.5 \text{ A})(0.05 \text{ m})}{2\pi} \ln\left(\frac{0.01 \text{ m} + 0.01 \text{ m}}{0.01 \text{ m}}\right) \\ &= 3.4657 \times 10^{-9} \text{ Wb} \end{aligned}$$

the unit of magnetic flux is called the weber and it is given by :

$$\text{Wb} = \text{T m}^2 = \frac{\text{m}^2}{\text{A}} \frac{\text{kg}}{\text{s}^2}$$

We know now how to calculate magnetic flux, but you should expect that we can do something with this flux to simplify problems. And your expectation would be right. We used electric flux in Gauss' law. We will use magnetic flux to find the induced emf. An induced emf can create a current, and this is the basic idea behind a generator. The law that governs this relationship between induced emf and magnetic flux is called *Faraday's law* after the scientist that discovered it. We will study this law in our next lecture.

Basic Equations

44 Faraday and Lenz

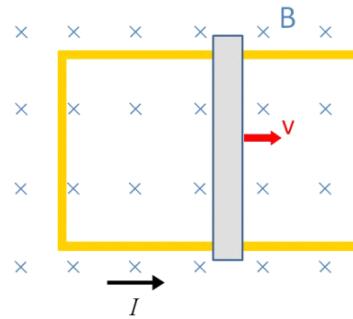
Fundamental Concepts

We talked about an induced electric field created by a magnetic field last lecture. We want to formalize that relationship in this lecture. Let's go back to our motional emf problem.

Question 223.44.1

Question 223.44.2

Question 223.44.3



We have a sliding bar, and a u-shaped conductor and a magnetic field. The moving bar makes the current flow. But now we know another way to express this. We can see that there is a magnetic flux through the loop consisting of the u-shaped conductor and the sliding bar. This flux going through the loop is changing. The area is getting larger, so the amount of field going through the loop is increasing. We can say the induced current is due to the changing loop area in the presence of the magnetic field, or a changing magnetic flux.

An important thing we learned is that the moving bar feels a resistive force due to the current and magnetic field. It seems like the magnetic field and current are resisting any change in our set up. We will see in this lecture that this is true in general.

It turns out that there is more than one way to cause an induced current. Any change in the magnetic flux is found to make a current flow. Remember in class we found that putting a magnet into or pulling the magnet out of a coil makes a current. In this case, the strength of the magnetic field changes, so the flux changes. Really any change in

magnetic flux makes a current flow.

Fundamental Concepts in the Lecture

- Changing magnetic flux makes an electric field—which has an associated potential difference or emf.
- The current caused by the induced emf travels in the direction that creates a magnetic field with flux opposing the change in the original flux through the circuit.
- The emf (potential difference) generated by a changing magnetic field is given by

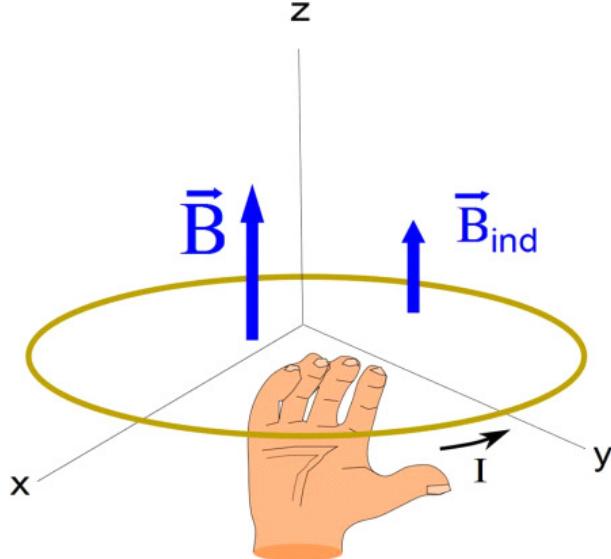
$$\mathcal{E} = -N \frac{\Delta \Phi_B}{\Delta t}$$

Lenz

What we are saying is that if we change the magnetic flux through a loop, we will get a current. The direction of current flow is not obvious. Lenz experimentally determined which way it will go. Here is his rule

The current caused by the induced emf travels in the direction that creates a magnetic field with flux opposing the change in the original flux through the circuit.

This takes a moment to digest. Let's take an example



Consider the case shown in the picture. Suppose the B -field gets smaller in time. If that is the case, then the induced current will try to keep the same number of field lines

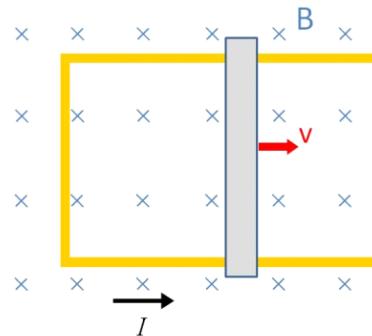
going through the loop. To do this, it will have to add field lines, because our field that is getting smaller will have fewer and fewer field lines. So in this case, the induced field $\tilde{\mathbf{B}}_{ind}$ will be in the same direction as $\tilde{\mathbf{B}}$ to try to keep the number of field lines the same. We find the current using our current-carrying wire right hand rule for magnetism. We imagine grabbing the wire such that our fingers curl into the loop the way $\tilde{\mathbf{B}}_{ind}$ goes through the loop. Then our thumb is in the direction of the current.

Question 223.44.4 -
223.44.11

Faraday

In our motional emf problem, the sliding bar in the magnetic field creates a potential difference, ΔV . It becomes an emf. We can use the symbol \mathcal{E} for our emf.

But then in considering Lenz's law, it was experimentally found that any change in flux causes a current. Then any change in flux must create an emf.



In this case the area is getting larger, and so the flux is getting larger. The induced current will oppose the change. So the induced magnetic field should go up through the center of the loop. Imagine sticking your fingers through the loop out of the page, then grabbing the loop (fingers still out of the page in the inside of the loop). Anywhere you grab the wire, your thumb is in the induced current direction.

Faraday's law of Magnetic Induction

Faraday wrote an equation to describe the emf that was given by changing a B -field. It combines what we know about magnetic flux and current from Lenz's law. Faraday did not know the source of the emf, it is a purely empirical equation. Here it is

$$\mathcal{E} = -N \frac{\Delta \Phi_B}{\Delta t} \quad (44.1)$$

The N is the number of turns in the coil (remember he used a coil, not just one loop). $d\Phi_B$ is the change in the magnetic flux. Our definition of magnetic flux is

$$\Phi_B = \int \vec{B} \cdot d\vec{A}$$

but for simple open surfaces we can gain some insight by writing the flux as

$$\Phi_B = BA \cos \theta$$

Then the induced emf would be given by

$$\mathcal{E} = -N \frac{\Delta \Phi}{\Delta t} \quad (44.2)$$

$$= -N \frac{(B_2 A_2 \cos \theta_2 - B_1 A_1 \cos \theta_1)}{\Delta t} \quad (44.3)$$

and we see that we get an emf if B , A , or θ change. We can write this as a differential if we let Δt get very small.

$$\mathcal{E} = -N \frac{d\Phi_B}{dt} \quad (44.4)$$

Suppose we have a simple flux $\Phi_B = \vec{B} \cdot \vec{A}$, then for this simple case

$$\begin{aligned} \mathcal{E} &= -N \frac{d}{dt} (\vec{B} \cdot \vec{A}) \\ &= -N \left(\vec{B} \cdot \frac{d}{dt} \vec{A} + \vec{A} \cdot \frac{d}{dt} \vec{B} \right) \end{aligned}$$

The first term shows our motional emf case. The area is changing in time. But the second term shows that if the field changes, we get an emf. This is the moving magnet in the coil case.

There are some great applications of induced emfs, from another design for circuit breakers to electric guitar pickups!

Question 223.44.12

- Question

223.44.17

Return to Lenz's law

Remember that Lenz's law says the current caused by the induced emf travels in the direction that creates a magnetic field with flux opposing the change in the original flux through the circuit. What if the current went the other way?

If that happened, then we could set up our bar on the rails, and give it a push to the right. With the current going down instead of up (for positive charge carriers) then we

would have a force on our bar-like segment of wire

$$F_I = BIL \sin \phi$$

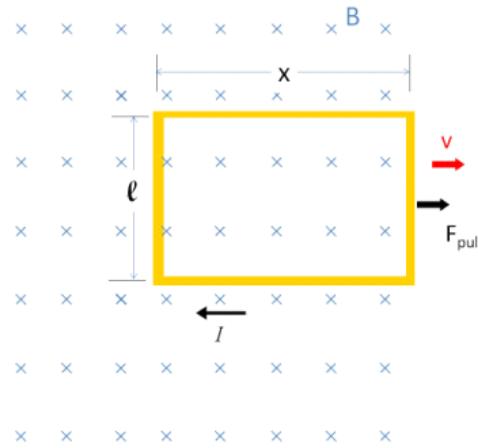
here $\sin \phi = 1$ so

$$F_I = BIL$$

It will be directed to the right. So the bar would accelerate to the right. That would increase the size of the loop, increasing the current. That would increase the force to the right, and our bar would soon zip off at amazing speed. But that does not happen. It would take ever more energy to make the bar go faster, with no input energy. So this would violate conservation of energy. Really Lenz's law just gives us conservation of energy again.

Pulling a loop from a magnetic field.

Let's try a problem. Suppose we have a wire loop. The loop is rectangular, with side lengths ℓ and x . Further suppose that the loop is in a region with magnetic field, but that it is on the edge of that field, so that if we pull it to the right, it will leave the field.



let's see if we can find the induced emf and current.

The Magnetic flux through the loop is changing. We can find an expression for the flux

$$\Phi_B = \vec{B} \cdot \vec{A}$$

or in this case

$$\Phi_B = B\ell x$$

We know the emf from Faraday's law

$$\mathcal{E} = -N \frac{d\Phi_B}{dt}$$

then

$$\mathcal{E} = -(1) \frac{d}{dt} (B\ell x)$$

The field is not changing strength, and the length ℓ is not changing. But along the x side, we are losing field. Remember that A in our flux equation is the area that actually has field and we have less area that has field all the time. We can see that

$$\mathcal{E} = -(1) \frac{d}{dt} (B\ell x) = -B\ell \frac{dx}{dt} = -B\ell v$$

where v is the speed at which we are pulling the wire loop. That is the speed at which our flux changes.

We can use Ohm's law to find the current,

$$I = \frac{\Delta V}{R} = \frac{\mathcal{E}}{R}$$

or

$$I = \frac{B\ell v}{R}$$

We could ask, how much work does it take to pull the wire out of the field? This is like our capacitor problem where we pulled a dielectric out of the middle of the capacitor.

The net force on the loop is not zero, because the field is no longer uniform. The right hand side of the loop is outside the field, and the left hand side is not. Of course, the top and bottom of the loop have opposite forces that balance each other. So the net force is due to the left hand side of the loop. Recall that

$$\vec{F} = I \vec{L} \times \vec{B}$$

We can see that in this case I is upward, and B is into the page. So there is a force to the left resisting our change flux. We must pull to overcome this force. The magnitude of this force is

$$F = I\ell B$$

and we know I so

$$F = \frac{B\ell v}{R} \ell B = \frac{B^2 \ell^2 v}{R}$$

Now we need to find the work done.

$$W = \int F dx$$

or, since our force will be constant until the loop leaves the magnetic field entirely,

$$W = F \int dx$$

which is not a hard integral to do. But instead of performing the integral, let's look at the integrand.

$$dW = F dx$$

if we divide both sides of our equation by dt we have

$$\frac{dW}{dt} = F \frac{dx}{dt}$$

we know that $P = dW/dt$ and $\frac{dx}{dt} = v$ and so we can write our equation as

$$\begin{aligned} P &= Fv \\ &= \frac{B^2 \ell^2 v^2}{R} \end{aligned}$$

which is how much power the magnetic field force provides in resisting. We must provide and equal power to move the loop. It will take time

$$\Delta t = \frac{\Delta x}{v}$$

to pull the loop a distance Δx . If we define our coordinates such that $x_i = 0$ then to pull out the loop, we will write this time as

$$\Delta t = \frac{x}{v}$$

so the work is

$$\begin{aligned} W &= P \Delta t \\ &= \frac{P}{R} \frac{B^2 \ell^2 v^2}{v} x \\ &= \frac{B^2 \ell^2 xv}{R} \end{aligned}$$

Incidentally, we learned from our demonstrations that induced currents can take energy out of a system, creating heat energy. From Ohm's law the power lost due to resistive heating would be

$$\begin{aligned} P &= I^2 R \\ &= \left(\frac{Blv}{R} \right)^2 R \\ &= \frac{B^2 l^2 v^2}{R} \end{aligned}$$

which is just the power we had to provide to make our loop move. So our work has moved the loop and heated up the wire.

We have created a current in a wire. This is the first step in building a generator. It cost us work to do this. In the next lecture, we will tackle more practical design and build generators and transformers. Then we will pause to think philosophically about what it means that a changing magnetic flux creates an electric field.

Basic Equations

45 Induced Fields

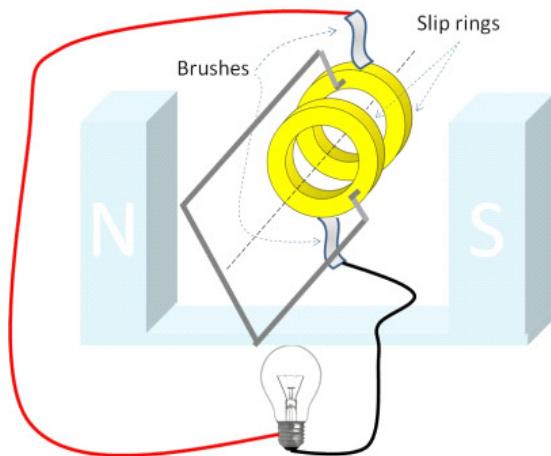
Fundamental Concepts

- Changing the commutator for slip rings makes a motor into a Generators
- Using alternating current, we can build an inductive device that can change from one voltage to another. This device is called a transformer.
- A more general form of Faraday's law is $\int \mathbf{E} \cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$

Generators

Question 223.45.1

Whether you are just plugging in an appliance, or preparing for an emergency, you likely would think of a generator as a source of electrical energy. Our studies so far have strongly hinted on how we would build an electric generator. In this lecture, we will fill in the details.



We can learn a lot by studying this device as an example. The figure shows the important parts of the generator (and a light bulb, which is not an important part of a generator, but just represents some device that will use the electrical current we make).

Question 223.45.2

The generator has at least one magnet. In the figure, there is one with a north end on the left and a south end on the right. A generator also has a wire loop. Usually in real generators, there are thousands of turns of wire forming the loop. In our picture, there is just one. The wire loop is connected to two metal rings. The rings will turn as the loop turns. Metal contacts (brushes) that can slip along the rings, but maintain an electrical connection, are placed on the rings. So as the rings turn, current can still flow through the connected wires (to the light bulb in this case).

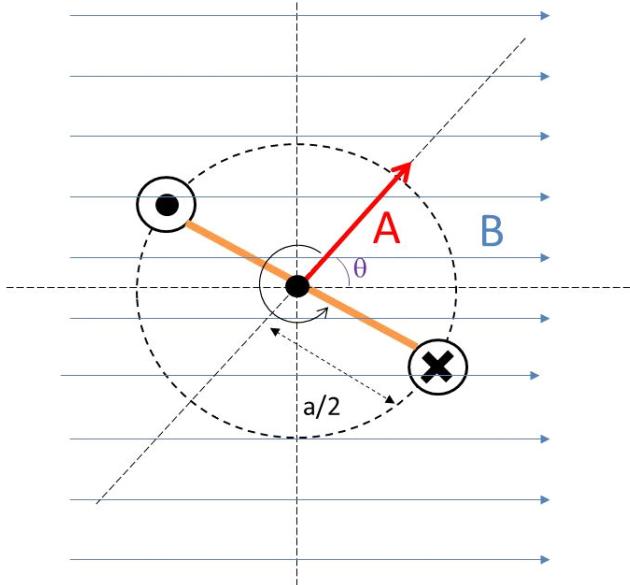
This should look familiar. This is the same basic setup as the motor, with a few exceptions. An important exception is that the commutator has been replaced by the set of rings. We will call these ring contacts *slip rings* because the wires can slip along them while still maintaining electrical contact because of the brushes. We have a current loop in a (nearly) uniform, constant field. If I look from the slip ring side of the loop, I have the same geometry we had before when we considered motors. This time I want to consider doing work to turn the loop, and find the induced emf in the loop. We start with Faraday's law

$$\mathcal{E} = -N \frac{d\Phi_B}{dt} \quad (45.1)$$

since in our special case we only have one loop, this is just

$$\mathcal{E} = -\frac{d\Phi_B}{dt} \quad (45.2)$$

Here is a the view looking at the cross section of the loop facing toward the slip rings.



Let's consider the flux through the loop. The definition we have for flux is

$$\begin{aligned}\Phi_B &= \mathbf{B} \cdot \mathbf{A} \\ &= BA \cos \theta \\ &= BA_{proj}\end{aligned}$$

where θ is the angle between the loop area vector and the magnetic field direction.

I want to write the flux in terms of the lengths of the wire. When the loop is standing up straight along the y -direction the projected area is just the area

$$A = \ell a$$

Then the projected area is

$$A_{proj} = \ell a \cos \theta$$

Let's check to make sure this works. When the loop is standing up straight along the y -direction $\theta = 0^\circ$, and $\cos \theta = 1$ so

$$A_{proj \ max} = \ell a \cos \theta = \ell a$$

so this works.

To find the emf generated, we need

$$\mathcal{E} = -\frac{d\Phi_B}{dt}$$

and only the area is changing, so

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -B \frac{dA_{proj}}{dt}$$

We realize that θ must change in time. We remember from Dynamics or PH121 that we can use $\theta = \omega t$ where ω is the angular speed of the rotating loop. Then

$$A_{proj} = \ell a \cos \omega t$$

and

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -B \frac{d}{dt} \ell a \cos \omega t$$

We recognize that θ changes as the loop turns Since B is not changing, the change in flux per unit time is just B times the change in area with time.

$$\mathcal{E} = B \ell a \omega \sin(\omega t)$$

Look at what we got! it is a sinusoidal emf. This will make a sinusoidal current!

$$\begin{aligned}I &= \frac{\mathcal{E}}{R} \\ &= \frac{B \ell a \omega \sin(\omega t)}{R}\end{aligned}$$

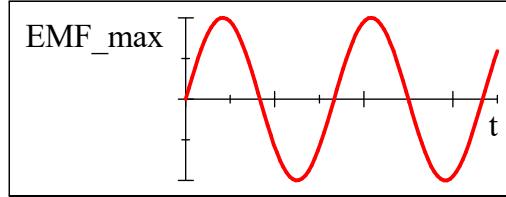
for a circuit. Our emf looks like

$$\mathcal{E} = \mathcal{E}_{max} \sin(\omega t) \tag{45.3}$$

where

$$\mathcal{E}_{\max} = Bl\omega \quad (45.4)$$

Here is a plot of the function



Of course this sinusoidal emf will create what we call an *alternating current*. This is how the current in the outlets in your house is generated.

Of course, our generator only has one coil. Actual generators have multiple coils.

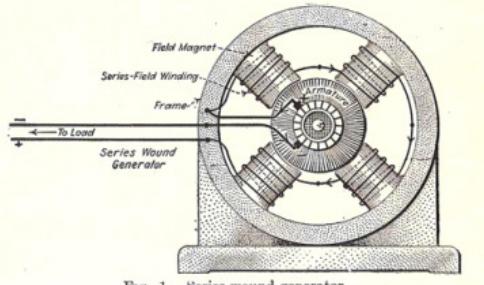
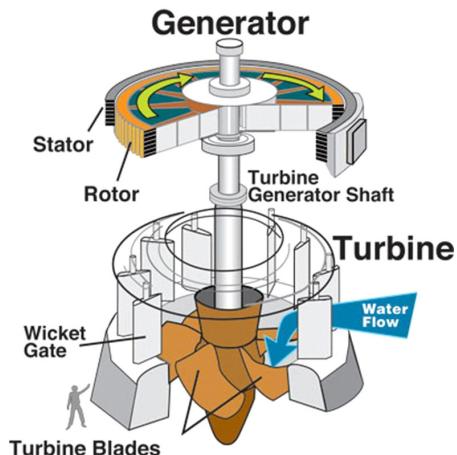


FIG. 1.—Series-wound generator.

Double Armature Generator (Public Domain Image)
and we need a source of work to turn the generator. A water turbine is an example,

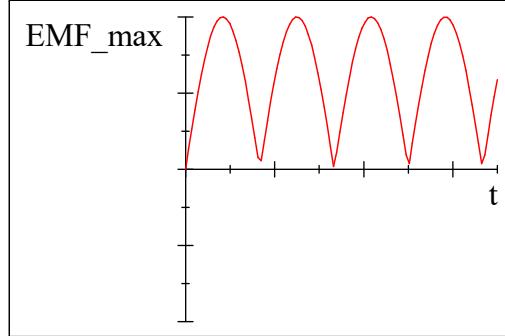


Water Turbine driven Generator (Public Domain Image courtesy U.S. Army Corps of Engineers)

or for emergencies, you might have a gasoline powered generator, or in a nuclear reactor you might have a steam driven generator.

DC current from a generator

We can also make a non-alternating current with a generator, but we have to get tricky to do it. We use the same idea we used to make a motor. We cut slots in the slip rings, so the current will switch directions every half turn. We get a kind of poor quality current from this because the emf still varies a lot.



Clever engineers design generators for non-alternating or *direct current* generators by overlapping several current loops at different angles. Each loop has its own cut slip rings. The combined currents smooth out the ripples we see in the previous figure. For semiconductor devices, special circuits are used to make the current very smooth.

Back emf

We can see that a motor is a DC generator run backwards. I just want to mention that when we talk about motors, we have to realize that as we send current into the motor coils, there will be an induced emf that will try to maintain the existing flux as the motor's loops turn. This emf will be in the opposite direction of the applied current! So it reduces the amount of work the motor can do. This is like the resistive force we encountered when we pulled a loop from a magnetic field last lecture. This resistive force is called the *back emf* and must be accounted for in motor design.

rms voltage

We can realize that we have a slight problem in talking about alternating voltages. The

voltage constantly changes. How do we describe what the voltage is?

We could give the max voltage—the amplitude of our $\mathcal{E}(t)$ curve. But the voltage is at the max only a small percentage of the time. We can't take the average. That is zero. And zero really doesn't describe our voltage well!.

The average doesn't work because our generators make the emf go negative. We could fix this by squaring the emf before we average it

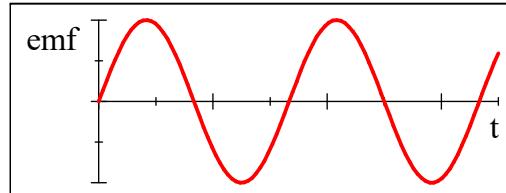
$$\begin{aligned}\overline{\mathcal{E}(t)} &= 0 \\ \overline{\mathcal{E}^2(t)} &\neq 0\end{aligned}$$

and this could work. But then we have the average voltage squared, and we really want just the voltage. No problem, let's take a square root.

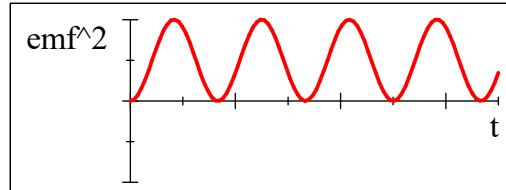
$$\sqrt{\overline{\mathcal{E}^2(t)}}$$

This has units of volts, is like an average of the emf, but doesn't cancel out because $\mathcal{E}(t)$ goes negative. Here is the process graphically.

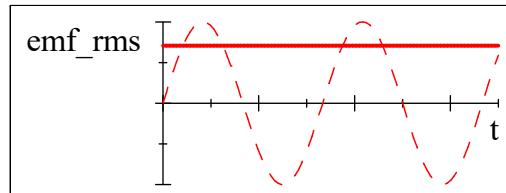
first $\mathcal{E}(t)$



now $\mathcal{E}^2(t)$



and finally $\sqrt{\overline{\mathcal{E}^2(t)}}$



What we did is take the square Root of the Mean of the Square of the emf. We can call this process the root-mean-square process or rms for short. This is more useful than the

mean voltage—which is zero for alternating voltages. It is a better estimate of the overall potential than the peak voltage value. So we often use rms voltages to describe sources of alternating current.

We can come up with a convenient way to find the rms emf. Consider that our alternating emf is given by

$$\mathcal{E} = \mathcal{E}_{\max} \sin(\omega t)$$

but we squared this

$$\mathcal{E}^2 = \mathcal{E}_{\max}^2 \sin^2(\omega t)$$

and then took an average. Suppose we average $\sin^2(\omega t)$ over a long time so that ωt gets large. We could say that And we want the case where θ_{\max} is large enough, our alternating voltage make many cycles. We would have would be

$$\begin{aligned}\overline{\mathcal{E}^2} &= \frac{1}{\Delta t} \int_{t_i}^{t_f} \mathcal{E}_{\max}^2 \sin^2(\omega t) dt \\ &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \int_{t_i}^{t_f} \sin^2(\omega t) dt\end{aligned}$$

but we have run into the integral of sine squared before. In equation 10.8 we found that

$$\int_{\text{many T}} \sin^2\left(\frac{k(r_2 + r_1)}{2} - \omega t + \phi_o\right) dt = \frac{1}{2}$$

and really it didn't matter much what the argument of \sin^2 was so long as we integrated over many periods. The integral is always 1/2. We can see this is true in our case by using a trig identity

$$\begin{aligned}\sin^2(\theta) &= \frac{1}{2}(1 - \cos(2\theta)) \\ \overline{\mathcal{E}^2} &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \int_{t_i}^{t_f} \sin^2(\omega t) dt \\ &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \int_{t_i}^{t_f} \left(\frac{1}{2} - \frac{1}{2} \cos 2t\omega\right) dt \\ &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \left(\int_{t_i}^{t_f} \frac{1}{2} dt - \int_{t_i}^{t_f} \frac{1}{2} \cos 2t\omega dt\right) \\ &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \left(\frac{1}{2}\Delta t - \int_{t_i}^{t_f} \frac{1}{2} \cos 2t\omega dt\right)\end{aligned}$$

The second integral over $\cos(2t\omega)$ will be zero. So we have

$$\begin{aligned}\overline{\mathcal{E}^2} &= \frac{\mathcal{E}_{\max}^2}{\Delta t} \frac{1}{2} \Delta t \\ &= \frac{1}{2} \mathcal{E}_{\max}^2\end{aligned}$$

Now we need to take a square root to finish the rms process.

$$\begin{aligned}\mathcal{E}_{rms} &= \sqrt{\bar{\mathcal{E}}^2} \\ &= \sqrt{\frac{1}{2}\mathcal{E}_{max}^2} \\ &= \frac{1}{\sqrt{2}}\mathcal{E}_{max}\end{aligned}$$

So the rms emf can be found from the max emf by dividing by the square root of 2.

This gives a pretty good idea of the nature of the voltage for alternating current. For example an *rms* emf value of 120 V would have a peak emf of

$$\begin{aligned}\mathcal{E}_{max} &= \sqrt{2}\mathcal{E}_{rms} \\ &= \sqrt{2}(120 \text{ V}) \\ &= 169.71 \text{ V}\end{aligned}$$

The *rms* value isn't the peak, it isn't the average (0) but it gives us a measure of how much voltage (and risk) we have.

We could find an *rms* current by using Ohm's law

$$\Delta V = IR$$

This would be true for

$$\mathcal{E}_{max} = I_{max}R$$

then we could find

$$I_{max} = \frac{\mathcal{E}_{max}}{R}$$

then if we divide by $\sqrt{2}$ we have

$$\frac{I_{max}}{\sqrt{2}} = \frac{\mathcal{E}_{max}}{\sqrt{2}R}$$

The right hand side is just

$$\frac{I_{max}}{\sqrt{2}} = \frac{\mathcal{E}_{rms}}{R}$$

so let's take

$$I_{rms} = \frac{I_{max}}{\sqrt{2}}$$

then Ohm's law for alternating currents becomes

$$I_{rms} = \frac{\mathcal{E}_{rms}}{R}$$

or

$$\mathcal{E}_{rms} = I_{rms}R$$

Transformers

Question 223.45.3

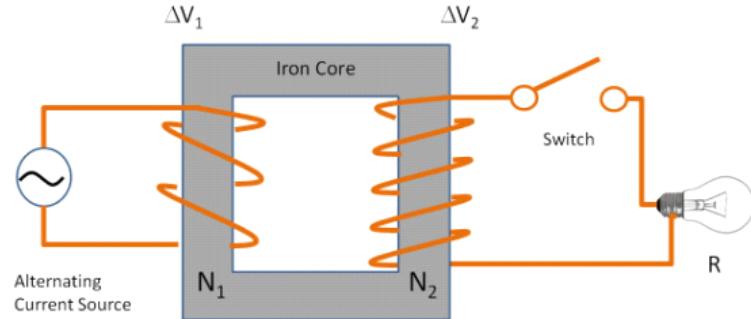
The power comes into our houses at about 120V. Your iPhone probably requires 3 V

to 5 V. How do we get the voltage we want out of what the power company delivers? You know the answer is to plug in your phone using a special adaptor. Lets see how it works.

Let's consider Faraday's law again. We know that

$$\mathcal{E} = \Delta V(t) = -N \frac{\Delta \Phi}{\Delta t}$$

Suppose we use Faraday's idea and hook two coils up next to each other.



One side we will hook to an alternating emf. We will call this side coil 1. The other side we will hook a second coil with some resistive load like a light bulb. We will call this coil 2. The iron core keeps the magnetic field inside, so the flux through coil 1 ends up going through coil 2. (think of all the little domains in the iron lining up along the field lines, and enhancing the field lines with their own induced fields).

The alternating potential from the source will create a change in flux in coil 1.

$$\mathcal{E}_1(t) = -N_1 \frac{\Delta \Phi_1}{\Delta t}$$

If little flux is lost in the iron, then we will retrieve most of the flux in coil 2 and an emf will be induced in the resister (light bulb in our case).

$$\mathcal{E}_2(t) = -N_2 \frac{\Delta \Phi_2}{\Delta t}$$

we just convinced ourselves that

$$\frac{\Delta \Phi_1}{\Delta t} \approx \frac{\Delta \Phi_2}{\Delta t}$$

so we can solve each equation for the change in flux term, and set them equal.

$$\begin{aligned} \frac{\mathcal{E}_1(t)}{N_1} &= -\frac{\Delta \Phi_1}{\Delta t} \\ \frac{\mathcal{E}_2(t)}{N_2} &= -\frac{\Delta \Phi_2}{\Delta t} \end{aligned}$$

so we have

$$\frac{\mathcal{E}_1(t)}{N_1} = \frac{\mathcal{E}_2(t)}{N_2} \quad (45.5)$$

If we solve for $\mathcal{E}_2(t)$ we can find the emf in coil 2.

$$\frac{N_2}{N_1} \mathcal{E}_1(t) = \mathcal{E}_2(t) \quad (45.6)$$

Question 223.45.4

You have probably already guessed how we make \mathcal{E}_2 to be some emf amount we want. We take, say, our wall current that has a *rms* value of $\mathcal{E}_1 = 120$ V. We pass it through this device we have built. We design the device so that $\frac{N_2}{N_1} \mathcal{E}_1$ gives just the potential that we want for \mathcal{E}_2 . If we want a lower emf, say 12 V, then we make $\frac{N_2}{N_1} = 0.1$ so

$$\frac{N_2}{N_1} \mathcal{E}_1 = 0.1 (120 \text{ V}) = 12 \text{ V} \quad (45.7)$$

This is part of what the wall adaptor does. Usually wall adapters also have some circuitry to make the alternating current into direct current.

Note that there is a cost to doing this. The power must be the same on both sides (or a little less on side 2). So

$$\mathcal{P}_{av} = I_{1,rms} \mathcal{E}_{1,rms} = I_{2,rms} \mathcal{E}_{2,rms}$$

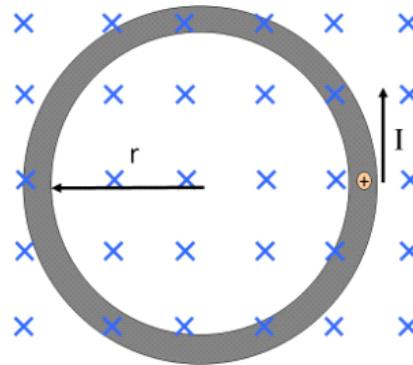
We can change the emf, but it will effect our current.

This device is called a transformer. Real transformers do lose power. Some loss is due to the fact that not all the *B*-field from coil 1 makes it inside coil 2. But real transformers are not too bad with efficiencies ranging from 90% to 99%.

Question 223.45.5

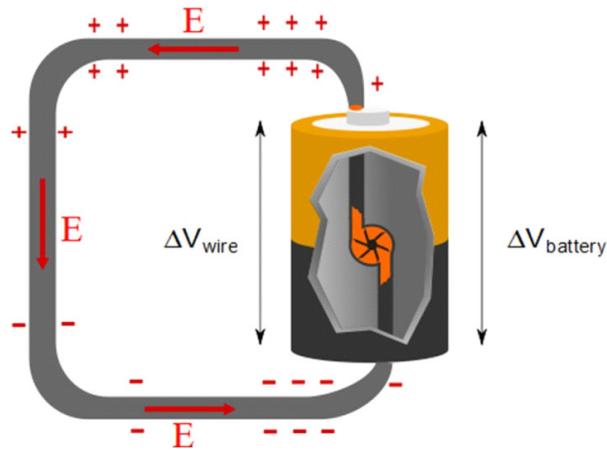
Induced Electric Fields

Consider again a magnetic field and a moving charge. If the field changes, the flux changes. Say, for example, that the field is increasing in strength.

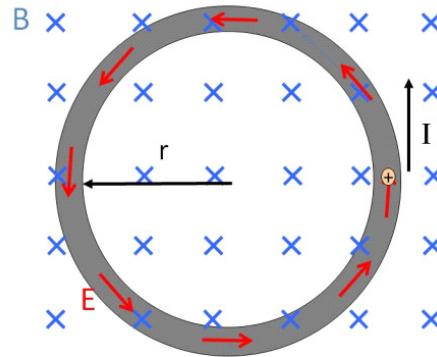


The charge will move in a circle within the wire. We now understand that this is because we have induced an emf. But think again about a battery.

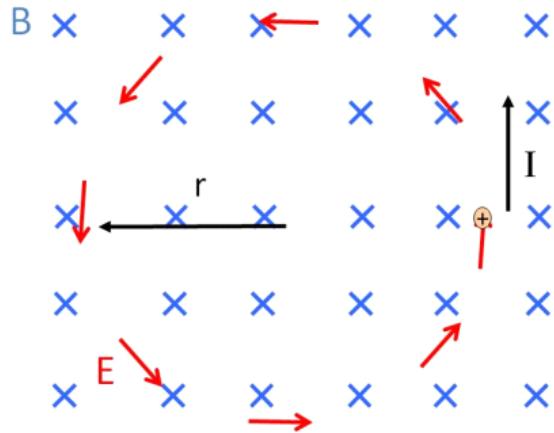
The battery makes an electric field inside a wire. Recall this figure



We must conclude that if we create an emf, we must have created an electric field.



This is really interesting. We now have a hint at how wireless chargers might work (we will return to this later). But now let's ask ourselves, do we need the wire there for this electric field to happen? Of course, the force on the charge is the same if there is no wire, so the E -field must be there whether or not there is a wire.

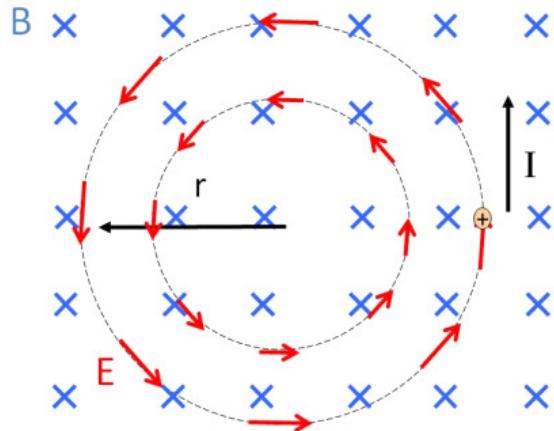


Question 223.45.6

In fact, the electric field is there in every place the magnetic field exists so long as the

Question 223.45.7

magnetic field continues to increase.



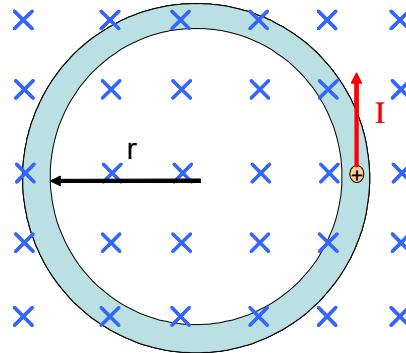
This is quite a profound statement. We have said that a changing magnetic field *creates* an electric field. Before, only charges could create electric fields, but in this case, the magnetic field is creating the electric field. Of course, we know that moving charges are making the magnetic field, so it is not totally surprising that the fields would be related.

This electric field is just like a field produced by charges in that it exerts a force

$$F = q_o E$$

on a charge q_o . But the electric field source is now very different.

Relationship between induced fields



Question 223.45.8

It would be nice to have a relationship between the changing B -field and the E -field that is created. It would be good to obtain the most general relationship we can that relates the electric field to the magnetic field. By understanding this relationship, we can hope to gain insight into how to build things, and into how the universe works. Let's start with a thought experiment.

Suppose we have a uniform but time varying magnetic field into the paper. In this field, we have a conducting ring. If the field strength is increasing, then the charges in the conducting loop shown will feel an induced emf, and they will form a current that is tangent to the ring.

Let's find the work required to move a charge once around the loop. The amount of potential energy difference is equal to the work done, so

$$|\Delta U| = |W|$$

but in terms of the electric potential this is

$$\Delta U = q\Delta V = q\mathcal{E}$$

so

$$|W| = |q\mathcal{E}|$$

Now let's do this another way. Let's use

$$W = \int \mathbf{F} \cdot d\mathbf{s}$$

The force making the current move is due to the induced potential difference. This is just

$$F = qE$$

which will not change as we go around the loop. The path will be along the loop, so

$$W = \int_{loop} F ds$$

and since the E -field is uniform in space at any given time as we travel around the loop,

$$W = F \int_{loop} ds = qE2\pi r$$

So we have two expressions for the work. Let's set them equal to each other

$$q\mathcal{E} = qE2\pi r$$

The field is then

$$\frac{\mathcal{E}}{2\pi r} = E \quad (45.8)$$

but

$$\mathcal{E} = -N \frac{d\Phi_B}{dt}$$

so

$$\begin{aligned} E &= \frac{-N}{2\pi r} \frac{d\Phi_B}{dt} \\ &= \frac{-1}{2\pi r} \frac{d\Phi_B}{dt} \end{aligned}$$

So if we know how our B -field varies in time, we can find the E -field. Let's rewrite this one more time

$$2\pi r E = -\frac{d\Phi_B}{dt}$$

Since the E -field is constant in as we go around the loop, we can recognize the LHS as

$$2\pi r E = \int \vec{E} \cdot d\vec{s}$$

which should be little surprise, since we found

$$\Delta V = \int \vec{E} \cdot d\vec{s}$$

to be our basic definition of the electric potential. So

$$\int \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt} \quad (45.9)$$

This is a more general form of Faraday's law of induction.

This electric field is fundamentally different than the E -fields we studied before. It is not a static field. If it were, then $\int \vec{E} \cdot d\vec{s}$ would be zero around a ring of current.

Think of conservation of energy. Around a closed loop $\Delta V = 0$ normally. Then

$$\Delta V = \int \vec{E} \cdot d\vec{s} = 0 \quad \text{no magnetic field}$$

But since $\int \vec{E} \cdot d\vec{s} \neq 0$ for our induced E -field, we must recognize that this field is different from those made by static charges. We call this field that does not return the charge to the same energy state on traversing the loop a *nonconservative field*. It is still just an electric field, but we are gaining energy from the magnetic field, so ΔV around

the loop is not zero.

The equation

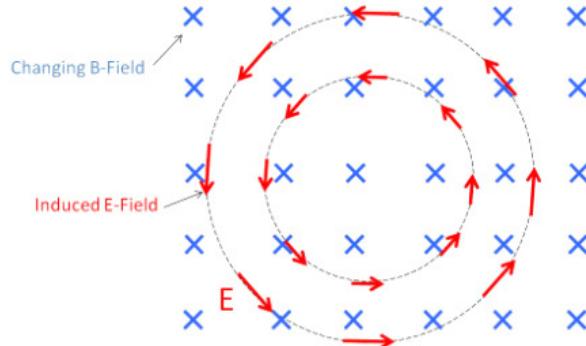
$$\int \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt} \quad (45.10)$$

is the most general form of Faraday's law, but it is hard to use in calculation for normal circuits where there is no magnetic field or where the fields are weak. So we won't use it as we design normal circuits (we will use the idea of inductance instead). But it plays a large part in the electromagnetic theory of optics (PH375). We will just get a taste of this here.

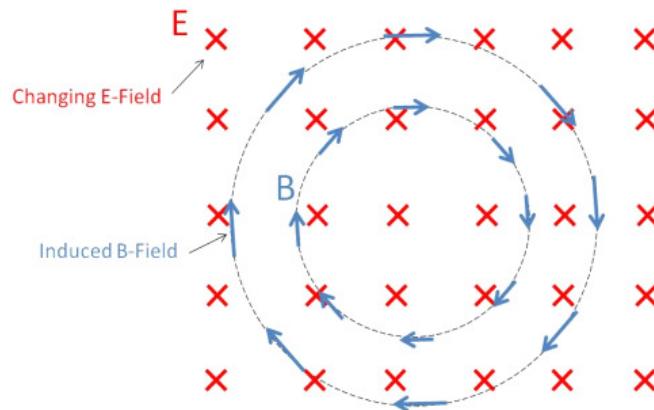
Electromagnetic waves

Question 223.45.9

Let's return to the idea that a changing magnetic field makes an electric field.



But what about a changing electric field?



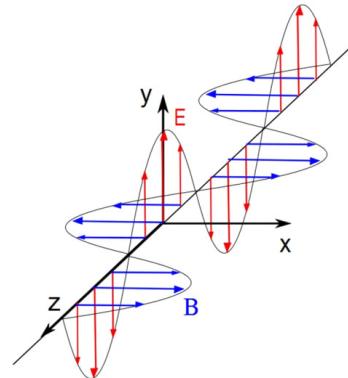
For the electric and magnetic field equations to be symmetric, the changing electric field must create a magnetic field. There is no requirement that the universe display

Question 223.45.10

such symmetry, but we have found that it usually does. Indeed, a changing electric field creates a magnetic field.

This foreshadows our final study of light. We learned earlier that light is an *electromagnetic* wave. What this means is that light is a wave in both the electric *and* magnetic fields.

Maxwell first predicted that such a wave could exist. The electric field of the wave changes in time like a sinusoid. But this change will produce a magnetic field that will also change in time. This changing magnetic field recreates the electric field—which recreates the magnetic field, etc. Thus the electromagnetic wave is *self-sustaining*. It can break off from the charges that create it and keep going forever because the electric field and magnetic field of the wave create each other. You often see the electromagnetic wave drawn like this:



Where you can see the electric and magnetic fields being created and recreated to make the wave self sustaining.

This is a direct result of Maxwell's study of electromagnetic field theory. Our more complete version of Faraday's law is one of the fundamental equations describing electromagnetic waves known as *Maxwell's Equations*.

$$\int \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

You might guess that the symmetry we have observed would give another similar equation relating the magnetic field and the electric flux.

$$\int \vec{B} \cdot d\vec{s} = +\frac{d\Phi_E}{dt}$$

and we will find that this is true! But we have yet to show that is so. Note that $\int \vec{B} \cdot d\vec{s}$ shows up in Ampere's law,

$$\int \vec{B} \cdot d\vec{s} = \mu_o I$$

so this last equation is not complete, but we are guessing that there is also the possibility of an induced magnetic field from a changing electric field, so we can predict that we need to modify Ampere's law to be

$$\int \vec{\mathbf{B}} \cdot d\vec{s} = \mu_o I + \frac{d\Phi_E}{dt}$$

but again we will have to show this later.

In the next lecture, we will take a break from this deep theoretical discussion, and learn how to use induction to make useful circuit devices that you used in ME210.

Basic Equations

46 Inductors

Fundamental Concepts

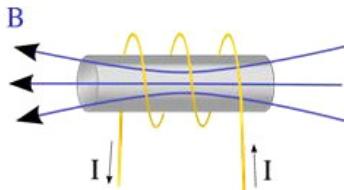
- The self inductance L has all the geometric and material properties of a coil or other inductor and it can be found using $L = N \frac{d\Phi_B}{dI}$
- The emf induced by an inductor is given by $\mathcal{E} \equiv -L \frac{\Delta I}{\Delta t}$
- For a solenoid, the inductance can be found to be $L = \mu_o n^2 V$
- The energy stored in the magnetic field is $U_L = \frac{1}{2} L I^2$ and the energy density in the magnetic field is $u_B = \frac{1}{2} \frac{1}{\mu_o} B^2$
- There is an *apparent* voltage drop across an inductor of $\Delta V_{L_{apparent}} = -L \frac{dI}{dt}$
- There is also a mutual inductance between two inductors given by $M_{12} = \frac{N_2 \Phi_{12}}{I_1}$

Self Inductance

Question 223.46.1

When we put capacitors and resistors in a circuit, we found that the current did not jump to its ultimate current value all at once. There was a time dependence. But really, even if we just have a resistor (and we always have some resistance) the current does not reach its full value instantaneously. Think of our circuits, they are current loops! So as the current starts to flow, Lenz's law tells us that there will be an induced emf that will oppose the flow. The potential drop across the resistor in a simple battery-resistor circuit is the potential drop due to the battery emf, *minus the induced emf*.

We can use this fact to control current in circuits. To see how, we can study a new case



Let's take a coil of wire wound around an iron cylindrical core. We start with a current as shown in the figure above. Using our right hand rule we can find the direction of the

B -field. But we now will allow the current to change. As it gets larger, we know

$$\mathcal{E} = -N \frac{d\Phi_B}{dt}$$

and we know that as the current changes, the magnitude of the B -field will change, so the flux through the coil will change. We will have an induced emf. We could derive this expression, but I think you can see that the induced emf is proportional to the *rate of change* of the current.

$$\mathcal{E} \equiv -L \frac{\Delta I}{\Delta t}$$

You might ask if the number of loops in the coil matters. The answer is—yes. Does the size and shape of the coil matter—yes. But we will include all these geometrical effects in the constant L called the *inductance*. It will hold all the material properties of the iron cored coil.

$$\mathcal{E} = -N \frac{d\Phi_B}{dt} \equiv -L \frac{dI}{dt}$$

so for this case

$$-N \frac{d\Phi_B}{dt} \frac{dt}{dI} \equiv -L$$

or

$$L = N \frac{d\Phi_B}{dI}$$

If we start with no current (so no flux), then our change in flux is the current flux minus zero. We can then say that

$$L = N \frac{\Phi_B}{I}$$

It might be more useful to write the inductance as

$$L = -\frac{\mathcal{E}_L}{\frac{dI}{dt}}$$

In designing circuits, we will usually just look up the inductance of the device we choose, like we looked up the resistance of resistors or the capacitance of the capacitors we use.

But for our special case of a simple coil, we can calculate the inductance, because we know the induced emf using Faraday's law

Inductance of a solenoid³⁵

Question 223.46.2

Let's extend our inductance calculation for a coil. Really the only easy case we can do

³⁵ Think of this like the special case of a capacitor made from two flat large plates, the parallel plate capacitor. It was somewhat ideal in the way we treated it. Our treatment of the special case of a coil will likewise be somewhat ideal.

is that of a solenoid (that's probably a hint for the test). So let's do it! We will just fill our solenoid with air instead of iron (if we have iron, we have to take into account the magnetization, so it is not terribly hard, but this is not what we want to concentrate on now). If the solenoid has N turns with length L and we assume that L is much bigger than the radius r of the loops then we can use our solution for the B -field created by a solenoid

$$\begin{aligned} B &= \mu_o n I \\ &= \mu_o \frac{N}{\ell} I \end{aligned}$$

The flux through each turn is then

$$\Phi_B = BA = \mu_o \frac{N}{\ell} IA$$

where A is the area of one of the solenoid loops. Then we use our equation for inductance for a coil

$$\begin{aligned} L &= N \frac{\Phi_B}{I} \\ &= N \frac{(\mu_o \frac{N}{\ell} IA)}{I} \\ &= \frac{(\mu_o N^2 A)}{\ell} \\ &= \frac{(\mu_o N^2 A) \ell}{\ell} \\ &= \frac{\mu_o N^2 A \ell}{\ell^2} \\ &= \frac{\mu_o N^2 V}{\ell^2} \\ &= \mu_o n^2 V \end{aligned}$$

where we used the fact that the volume of the solenoid is $V = A\ell$.

Many inductors built for use in electronics are just this, air filled solenoids. So this really is a somewhat practical solution.

Energy in a Magnetic Field

Question 223.46.3

An inductor, like a capacitor, stores energy in its field. We would like to know how much energy an inductor can store. From basic circuit theory we know the power in a circuit will be

$$\mathcal{P} = I\Delta V$$

If we just have an inductor, then the power removed from the circuit is

$$\begin{aligned}\mathcal{P}_{cir} &= I\Delta V = I\mathcal{E} \\ &= I \left(-L \frac{dI}{dt} \right) \\ &= -LI \frac{dI}{dt}\end{aligned}$$

As with a resistor, we are taking power *from the circuit* so the result is negative. But unlike a resistor, this power is not being dissipated as heat. It is going into the magnetic field of the inductor. Therefore, we expect the power stored in the inductor field to be

$$\mathcal{P}_L = -\mathcal{P}_{cir} = LI \frac{dI}{dt}$$

Power is the time rate of change of energy, so we can write this power delivered to the inductor as

$$\frac{dU_L}{dt} = LI \frac{dI}{dt}$$

Multiplying by dt gives

$$dU_L = LI dI$$

To find the total energy stored in the inductor we must integrate over I .

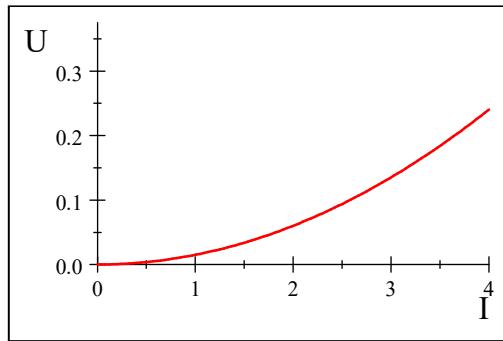
$$\begin{aligned}U_L &= \int dU_L \\ &= \int_0^I LI dI \\ &= L \int_0^I IdI \\ &= \frac{1}{2} LI^2\end{aligned}$$

Thus,

$$U_L = \frac{1}{2} LI^2$$

is the energy stored in the magnetic field of the inductor.

Suppose we have an inductor $L = 30.0 \times 10^{-3}$ H. Plotting shows us the dependence of U_L on I .



We should take a moment to see how our inductor compares to a capacitor as an energy storage device. The energy stored in the electric field of a capacitor

$$U_L = \frac{1}{2}L(I)^2$$

$$U_C = \frac{1}{2}C(\Delta V)^2$$

Notice that Remarkable similarity!

Energy Density in the magnetic field

Question 223.46.4

We found that there was energy stored in the electric field of a capacitor. Is the energy stored in the inductor really stored in the magnetic field of the inductor? We believe that this is just the case, the energy, U_L , is stored in the field. We would like to have an expression for the density of the energy in the field.

To see this, let's start with the inductance of a solenoid.

$$L = \mu_o n^2 A \ell$$

The magnetic field is given by

$$B = \mu_o n I$$

then the energy in the field is given by

$$\begin{aligned} U_B &= \frac{1}{2} L I^2 \\ &= \frac{1}{2} \mu_o n^2 A \ell I^2 \end{aligned}$$

If we rearrange this, we can see the solenoid field is found in the expression twice

$$\begin{aligned} U_B &= \frac{1}{2} (\mu_o n I) A \ell \frac{\mu_o n I}{\mu_o} \\ &= \frac{1}{2 \mu_o} B^2 A \ell \end{aligned}$$

and the energy density is

$$\begin{aligned} u_B &= \frac{U_B}{A \ell} \\ &= \frac{1}{2} \frac{1}{\mu_o} B^2 \end{aligned}$$

Just like our energy density for the electric field, we derived this for a specific case, a solenoid. But this expression is general. We should compare to the energy density in the electric field.

$$u_E = \frac{1}{2} \epsilon_0 E^2$$

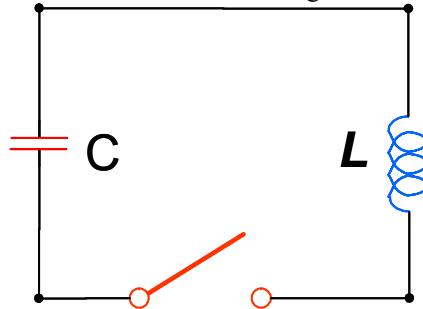
Again, note the similarity!

Oscillations in an LC Circuit

We introduce a new circuit symbol for inductors



It looks like a coil, for obvious reasons. We can place this new circuit element in a circuit. But what will it do? To investigate this, let's start with a simple case, a circuit with a charged capacitor and an inductor and nothing else.



Let us make two unrealistic assumptions (we will relax these assumptions later).

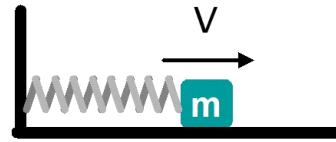
Assumption 1: There is no resistance in our LC circuit.

Assumption 2: There is no radiation emitted from the circuit.

Given these two assumptions, there is no mechanism for energy to escape the circuit.

Question 223.46.5 Energy must be conserved. Can we describe the charge on the capacitor, the current, and the energy as a function of time?

It may pay off to recall some details of oscillators. Energy of the Simple Harmonic Oscillator



Remember from Dynamics or PH121 that a mass-spring system will oscillate. The mass has kinetic energy because the mass is moving

$$K = \frac{1}{2}mv^2 \quad (46.1)$$

for our Simple Harmonic Oscillator we know that the position of the mass as a function of time is given by

$$x(t) = x_{\max} \cos(\omega t + \phi)$$

and the speed as a function of time is

$$v(t) = -\omega x_{\max} \sin(\omega t + \phi)$$

then the kinetic energy as a function of time is

$$\begin{aligned} K &= \frac{1}{2}m(-\omega x_{\max} \sin(\omega t + \phi))^2 \\ &= \frac{1}{2}m\omega^2 x_{\max}^2 \sin^2(\omega t + \phi) \\ &= \frac{1}{2}m\frac{k}{m}x_{\max}^2 \sin^2(\omega t + \phi) \\ &= \frac{1}{2}kx_{\max}^2 \sin^2(\omega t + \phi) \end{aligned}$$

The spring has potential energy given by

$$U = \frac{1}{2}kx^2 \quad (46.2)$$

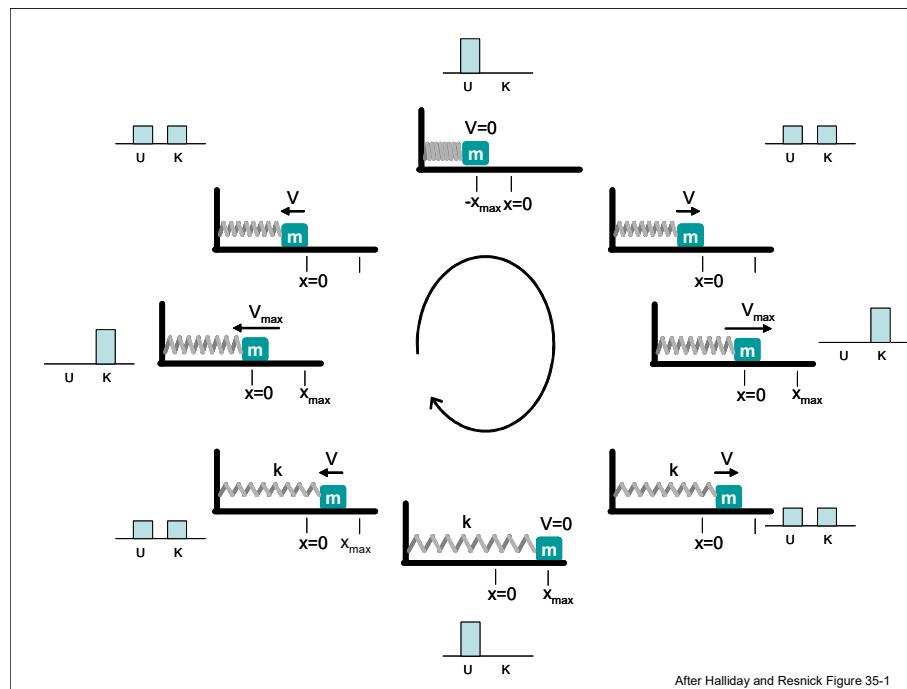
For our mechanical oscillator the potential as a function of time is

$$U = \frac{1}{2}kx_{\max}^2 \cos^2(\omega t + \phi)$$

The total energy is given by

$$\begin{aligned}
 E &= K + U \\
 &= \frac{1}{2} kx_{\max}^2 \sin^2(\omega t + \phi) + \frac{1}{2} kx_{\max}^2 \cos^2(\omega t + \phi) \\
 &= \frac{1}{2} kx_{\max}^2 (\sin^2(\omega t + \phi) + \cos^2(\omega t + \phi)) \\
 &= \frac{1}{2} kx_{\max}^2
 \end{aligned}$$

We can see that the total energy won't change, and the energy switches back and forth from kinetic to potential as the mass moves back and forth. If we plot the kinetic and potential energy at points along the mass' path we get something like this.

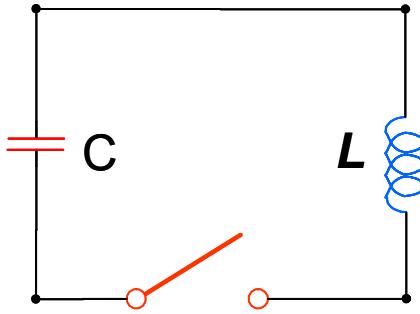


After Halliday and Resnick Figure 35-1

Question 223.46.6

One of the important uses of an inductor is to create *electrical oscillations*. Having recalled what oscillations look like, we can see that a LC circuit will have an oscillating current.

here is our circuit again.



We will start with the switch open the capacitor charged to its maximum value Q_{\max} . For $t > 0$ the switch is closed. Recall that the energy stored in the capacitor is

$$U_C = \frac{Q^2}{2C}$$

and the energy stored in the inductor is

$$U_L = \frac{1}{2}I^2L$$

The total energy (because of our assumptions) is

$$\begin{aligned} U &= U_C + U_L \\ &= \frac{Q^2}{2C} + \frac{1}{2}I^2L \end{aligned}$$

The change in energy over time must be zero (again because of our assumptions) so

$$\begin{aligned} \frac{dU}{dt} &= 0 \\ &= \frac{d}{dt} \left(\frac{Q^2}{2C} + \frac{1}{2}I^2L \right) \\ &= \frac{Q}{C} \frac{dQ}{dt} + LI \frac{dI}{dt} \end{aligned}$$

We recall that

$$I = \frac{dQ}{dt}$$

$$\begin{aligned} 0 &= \frac{Q}{C} \left(\frac{dQ}{dt} \right) + LI \frac{dI}{dt} \\ 0 &= \frac{Q}{C} (I) + LI \frac{dI}{dt} \\ 0 &= \frac{Q}{C} I + LI \frac{d \left(\frac{dQ}{dt} \right)}{dt} \\ 0 &= \frac{Q}{C} + L \frac{d^2Q}{dt^2} \end{aligned}$$

or

$$\frac{d^2Q}{dt^2} = -\frac{Q}{LC}$$

This is a differential equation that we recognize from M316. It looks just like the differential equation for oscillatory motion! We try a solution of the form

$$Q = A \cos(\omega t + \phi)$$

then

$$\frac{dQ}{dt} = -A\omega \sin(\omega t + \phi)$$

and

$$\frac{d^2Q}{dt^2} = -A\omega^2 \cos(\omega t + \phi)$$

thus

$$A\omega^2 \cos(\omega t + \phi) = -\frac{1}{LC} A \cos(\omega t + \phi)$$

This is indeed a solution if

$$\omega = \frac{1}{\sqrt{LC}}$$

When $\cos(\omega t + \phi) = 1$, $Q = Q_{\max}$, thus

$$Q = Q_{\max} \cos(\omega t + \phi)$$

Now recall,

$$\begin{aligned} I &= \frac{dQ}{dt} \\ &= \frac{d}{dt}(Q_{\max} \cos(\omega t + \phi)) \\ &= -\omega Q_{\max} \sin(\omega t + \phi) \end{aligned}$$

We would like to determine ϕ . We use the initial conditions $t = 0$, $I = 0$ and $Q = Q_{\max}$. Then

$$0 = -\omega Q_{\max} \sin(\phi)$$

This is true for $\phi = 0$. Then

$$\begin{aligned} Q &= Q_{\max} \cos(\omega t) \\ I &= -\omega Q_{\max} \sin(\omega t) \\ &= -I_{\max} \sin(\omega t) \end{aligned}$$

We can use the solution for the charge on the capacitor and the current in the inductor

as a function of time to expand our energy equation

$$\begin{aligned} U &= U_C + U_L \\ &= \frac{Q^2}{2C} + \frac{1}{2}I^2L \\ &= \frac{1}{2C}Q_{\max}^2 \cos^2(\omega t) + \frac{1}{2}LI_{\max}^2 \sin^2(\omega t) \end{aligned}$$

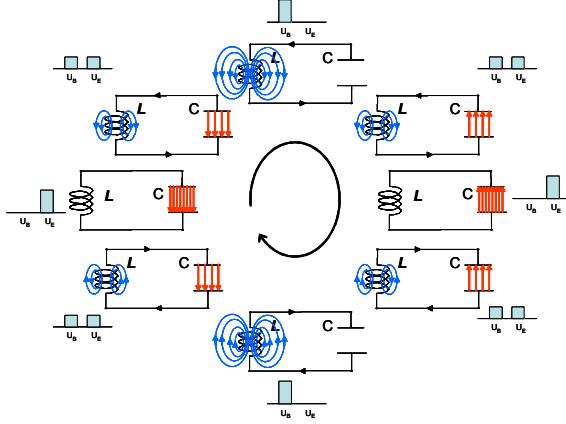
This looks a lot like our kinetic and potential energy equation for a mass-spring system. The energy shifts from the capacitor to the inductor and back like energy shifted from kinetic to potential energy for our mass-spring, with the components out of phase by 90° . By energy conservation, we know that

$$\frac{1}{2C}Q_{\max}^2 = \frac{1}{2}LI_{\max}^2$$

that is, the maximum energy in the capacitor equals the maximum energy in the inductor. Then the total energy

$$\begin{aligned} U &= \frac{1}{2C}Q_{\max}^2 \cos^2(\omega t) + \frac{1}{2}LI_{\max}^2 \sin^2(\omega t) \\ &= \frac{1}{2C}Q_{\max}^2 \cos^2(\omega t) + \frac{1}{2C}Q_{\max}^2 \sin^2(\omega t) \\ &= \frac{Q_{\max}^2}{2C} \end{aligned}$$

which must be the case if energy is conserved. We can plot the capacitor and inductor energies at points in time as the current switches back and forth.



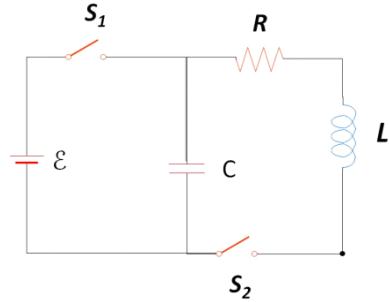
After Halliday and Resnick Figure 35-1

This is very much like our harmonic oscillator picture. We can see that we have, indeed made an electronic oscillator.

This type of circuit is a major component of radios which need a local oscillatory circuit to operate.

The RLC circuit

As fascinating as the last section was, we know there really is some resistance in the wire. So the restriction of no resistance needs to be relaxed in our analysis.



We can use the circuit in the picture to imagine an LRC circuit. At first, we will keep S_2 open and close S_1 to charge up the capacitor. Then we will close S_1 and open S_2 . What will happen?

It is easier to find the current and charge on the capacitor as a function of time by using energy arguments. The resistor will remove energy from the circuit by dissipation (getting hot). The circuit has energy

$$U = \frac{Q^2}{2C} + \frac{1}{2}LI^2 \quad (46.3)$$

so from the work energy theorem,

$$W_{nc} = \Delta U$$

the energy lost will be related to a change in the energy in the capacitor and the inductor. Let's look at the rate of energy loss again

$$\begin{aligned} \frac{dU}{dt} &= \frac{d}{dt} \left(\frac{Q^2}{2C} + \frac{1}{2}LI^2 \right) \\ &= \frac{Q}{C} \frac{dQ}{dt} + LI \frac{dI}{dt} \end{aligned} \quad (46.4)$$

but this must be equal to the loss rate. The power lost will be $P = I^2R$

$$-I^2R = \frac{Q}{C} \frac{dQ}{dt} + LI \frac{dI}{dt} \quad (46.5)$$

This is a differential equation we can solve, let's first rearrange, remembering that

$$I = \frac{dQ}{dt}$$

then

$$\begin{aligned}-I^2R &= \frac{Q}{C}I + LI\frac{dI}{dt} \\ -IR &= \frac{Q}{C} + L\frac{dI}{dt}\end{aligned}$$

again using $I = \frac{dQ}{dt}$

$$+L\frac{d^2Q}{dt^2} + \frac{dQ}{dt}R + \frac{Q}{C} = 0 \quad (46.6)$$

This is a good exercise for those of you who have taken math 316. This is just like the equation governing a damped harmonic oscillator. The solution is

$$Q = Q_{\max}e^{-\frac{Rt}{2L}} \cos \omega_d t \quad (46.7)$$

where the angular frequency, ω_d is given by

$$\omega_d = \left(\frac{1}{LC} - \left(\frac{R}{2L} \right)^2 \right)^{\frac{1}{2}} \quad (46.8)$$

Remember that for a damped harmonic oscillator

$$x(t) = Ae^{-\frac{b}{2m}t} \cos(\omega t + \phi)$$

and

$$\omega = \left(\frac{k}{m} - \left(\frac{b}{2m} \right)^2 \right)^{\frac{1}{2}}$$

The resistance acts like a damping coefficient! Suppose

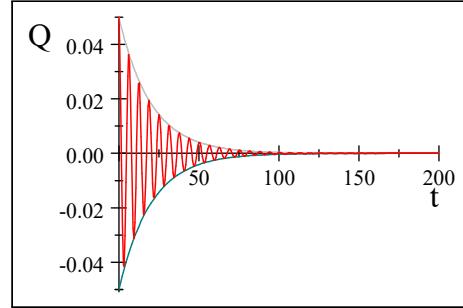
$$Q_{\max} = 0.05 \text{ C}$$

$$R = 5 \Omega$$

$$L = 50 \text{ H}$$

$$C = 0.02 \text{ F}$$

we have a graph that looks like this.



The gray lines are

$$\pm Q_{\max} e^{-\frac{Rt}{2L}} \quad (46.9)$$

They describe how the amplitude changes. We call this the *envelope* of the curve.

Let's look at

$$\omega_d = \left(\frac{1}{LC} - \left(\frac{R}{2L} \right)^2 \right)^{\frac{1}{2}} \quad (46.10)$$

If $\omega_d = 0$ then

$$\begin{aligned} 0 &= \frac{1}{LC} - \left(\frac{R}{2L} \right)^2 \\ \frac{1}{LC} &= \left(\frac{R}{2L} \right)^2 \\ 2L\sqrt{\frac{1}{LC}} &= R \end{aligned}$$

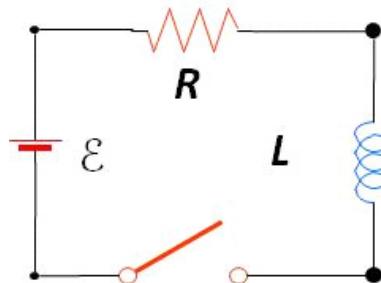
or

$$R = \sqrt{\frac{4L}{C}} \quad (46.11)$$

We know that if $\omega_d = 0$ there is no oscillation. We will call this the critical resistance, R_c . When the resistance is $R \geq R_c$ there will be no oscillation. These represent the cases of being critically damped ($R = R_c$) and overdamped ($R > R_c$). If $R < R_c$ we are underdamped, and the circuit will oscillate.

We don't know how to make electromagnetic waves yet, but we will in a few lectures. Those waves carry what we call radio signals. To make the waves, we often use circuits with resistors, capacitors, and inductors to provide the oscillation. You can guess that if Q on the capacitor oscillates, so does the current. This oscillating current is what we use to drive the radio antenna.

Now that we have some resistance, we could consider a circuit with just an inductor and a resistor and a battery.



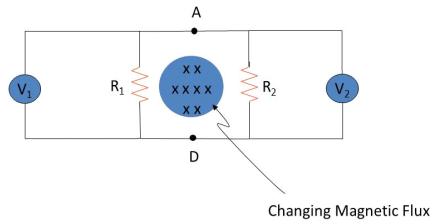
This is a little harder to deal with than it might appear. Let's examine the difficulties in thinking about such a circuit in the next section.

Return to Non-Conservative Fields

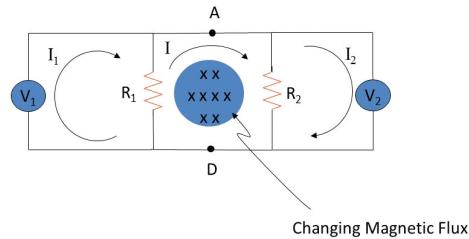
A few decades ago, we could have stopped here in an engineering class in considering an LRC circuit. But as electrical devices become ever more complicated, it might be good if we examine circuits with inductors and resistors more carefully. A few lectures ago we found that

$$\int \mathbf{E} \cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$

implies a non-conservative electric field. We should take a moment to see what this means. We should also investigate mutual inductance, which has become a major engineering technique for wireless power. First let's consider the following circuit.[?]



notice that there is no battery. If the field flux changes, will there be a potential difference measured by the voltmeters? Let's use conservation of energy to analyze the circuit. I can draw in guesses for the currents.



At the junction, we can use conservation of charge to see how the currents combine or divide. This will allow us to find the voltages.

But recall that

$$\oint \mathbf{E} \cdot d\mathbf{s} = 0$$

is a statement of conservation of energy. In electronics, we sometimes call this Kirchhoff's loop rule. And we learned that this is not true for induced emfs. So in the middle loop Kirchhoff's loop rule—conservation of energy—is not true! Some energy is transferred into or out of the circuit. We now know that is because of the changing magnetic field,

$$\oint \mathbf{E} \cdot d\mathbf{s} = \mathcal{E} = -\frac{d\Phi_B}{dt}$$

for the middle loop. In this case, \mathcal{E} comes just from the changing external flux. It does *not* depend on R_1 or on R_2 .

We can write a conservation of energy equation (per unit charge) for each loop.

$$\begin{aligned} I_1 R_i - IR_1 &= 0 \\ -IR_1 - IR_2 + \mathcal{E} &= 0 \\ I_2 R_i - IR_2 &= 0 \end{aligned}$$

where R_i is the internal resistance of the voltmeters. If there were no \mathcal{E} , then the voltmeters would not read anything, but now we see that

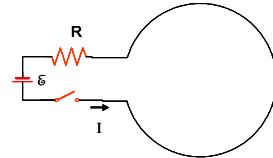
$$\begin{aligned} |V_1| &= I_1 R_i \approx IR_1 \\ |V_2| &= I_2 R_i \approx IR_2 \end{aligned}$$

This seems crazy. Each volt meter reads a different voltage.

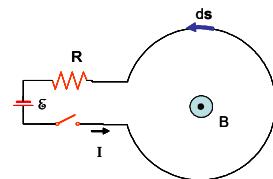
To understand this, remember that our induced field is not a conservative field. It is providing some energy. As we go around the loop we no longer expect to get back to our starting voltage. We have gained (or lost) some energy from the changing magnetic field. And for non-conservative fields, $\oint \mathbf{E} \cdot d\mathbf{s}$ is path dependent.

So as crazy as it seems, this is actually what we would find, each volt meter reads a different voltage.

To try to make this idea of inductance make some sense, let's take another strange circuit.



There is a battery, and resister, and a single loop inductor. When the switch is thrown, the current will flow as shown. The current will create a magnetic field that is out of the page in the center of the loop. Since the loop, itself, is creating this field, let's call this field a *self field*.



Consider this self-field for a moment. When we studied charge, we found that charge created an electric field. That electric field could make *another* charge accelerate. But

the electric field created by a charge does not make the charge that created it accelerate. This is an instance of a self-field, an electric self-field. Now with this background, let's return to our magnetic self-field.

Let's take Faraday's law and apply it to this circuit. Let me choose an area vector \mathbf{A} that is the area of the big loop and positive out of the page. Again, let's use conservation of energy (Kirchhoff's loop law). Let's find $\oint \mathbf{E} \cdot d\mathbf{s}$ for the entire circuit. We can start with the battery. Since there is an electric field inside the battery we will have a component of $\oint_{bat} \mathbf{E} \cdot d\mathbf{s}$ as we cross it. The battery field goes from positive to negative. If we go counter-clockwise, our $d\mathbf{s}$ direction traverses this from negative to positive, so the electric field is up and the $d\mathbf{s}$ direction is down, we have

$$\oint_{bat} \mathbf{E} \cdot d\mathbf{s} = -\mathcal{E}_{bat}$$

for this section of the circuit. Suppose we have ideal wires. If the wire has no resistance, then it takes no work to move the charges through the wire. In this case, an electron launched by the electric field in the battery just coasts from the battery to the resistor. There is no need to have an acceleration in the ideal wire. The electric potential won't change from the battery to the resistor. So there doesn't need to be a field in this ideal wire part to keep the charges going. But let's next we consider the resistor. There is a potential change as we go across it. And if there is a change in potential, there must be an electric field. So the resistor also has an electric field inside of it. We have a component of $\oint_R \mathbf{E} \cdot d\mathbf{s}$ that is equal to $\mathcal{E}_R = IR$ from this field.

$$\oint_R \mathbf{E} \cdot d\mathbf{s} = IR$$

Now we come to the big loop part. Since we have ideal wire, there is no resistance in this part so there is no voltage drop for this part of the circuit. All the energy that was given to the electrons by the battery was lost in the resistor. They just coast back to the other terminal of the battery. Since there is no voltage drop in the big loop,

$$\mathcal{E}_{\text{big loop}} = 0$$

there is no electric field in the big loop either. Along the big loop, $d\mathbf{s}$ is certainly not zero. so

$$\mathcal{E}_{\text{big loop}} = \oint_{\text{big loop}} \mathbf{E} \cdot d\mathbf{s} = 0$$

For the total loop we would have

$$\oint \mathbf{E} \cdot d\mathbf{s} = -\mathcal{E}_{batt} + IR + 0 \quad (46.12)$$

Normally, conservation of energy would tell us that all this must be zero, since the sum of the energy changes around the loop must be zero if no energy is lost. But now we

know energy *is* lost in making a magnetic field.

Consider the magnetic flux through the circuit. The magnetic field is made by the current in the circuit. Note that we arranged the circuit so the battery and resistor are in a part that has very little area, so we can ignore the flux through that part of the circuit. Most of the flux will go through the big loop part. The magnetic field is out of the paper inside of the loop. The flux is

$$\Phi_B = \oint \mathbf{B} \cdot d\mathbf{A} \quad (46.13)$$

and \mathbf{B} and \mathbf{A} are in the same direction. Φ_B is positive.

Then from Biot-Savart

$$\mathbf{B} = \frac{\mu_0 I}{4\pi} \oint \frac{d\mathbf{s} \times \hat{\mathbf{r}}}{r^2} \quad (46.14)$$

Let's write this as

$$\begin{aligned} \mathbf{B} &= I \left(\frac{\mu_0}{4\pi} \oint \frac{d\mathbf{s} \times \hat{\mathbf{r}}}{r^2} \right) \\ &= I (\text{geometry factor}) \end{aligned} \quad (46.15)$$

If the geometry of the situation does not change, then B and I are proportional. Since $B \propto I$, then $\Phi_B \propto I$ since the integral in Biot-Savart is the surface integral of \mathbf{B} , and \mathbf{B} is everywhere proportional to I . Instead of using Biot-Savart, let's just define a constant of proportionality that will contain all the geometric factors. We could give it the symbol, L . Then

$$\Phi_B = LI \quad (46.16)$$

where L is my geometry factor. But we recognize this geometry factor. It is just our inductance! This is what inductance is. It is all the geometry factors that make up our loop that will make the magnetic field if we put a current through it.

Assuming I don't change the geometry, then the inductance won't change and we have

$$\frac{d\Phi_B}{dt} = L \frac{dI}{dt} \quad (46.17)$$

and Faraday's law gives us

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -L \frac{dI}{dt} \quad (46.18)$$

Which says that we should not have expected $\oint \mathbf{E} \cdot d\mathbf{s} = 0$ for our case as we traverse the entire circuit. Integrating $\oint \mathbf{E} \cdot d\mathbf{s}$ around the whole circuit including the big loop should not bring us back to zero voltage. We have lost energy in making the field.

Instead it gives

$$\oint \mathbf{E} \cdot d\mathbf{s} = -L \frac{dI}{dt}$$

We are dealing with non-conservative fields. So we have some energy loss like we

would with a frictional force. It took some energy to make the magnetic field!

With this insight, we can now make a new statement of conservation of energy for such a situation. Integrating around the whole circuit gives

$$\oint \mathbf{E} \cdot d\mathbf{s} = -\mathcal{E}_{bat} + \mathcal{E}_R$$

Which we now realize should give $-L\frac{dI}{dt}$ so

$$\oint \mathbf{E} \cdot d\mathbf{s} = -\mathcal{E}_{bat} + \mathcal{E}_R = -L\frac{dI}{dt}$$

or more succinctly

$$-\mathcal{E}_{batt} + IR = -L\frac{dI}{dt}$$

Now I can take the RHS to the left and find

$$\mathcal{E}_{batt} - IR - L\frac{dI}{dt} = 0 \quad (46.19)$$

which accounts for all of the energy in the situation, so now we see that energy is conserved. For those of you who go on in your study of electronics. this looks like a Kirchhoff's rule with $-L\frac{dI}{dt}$ being a voltage drop across the single loop inductor. Under most conditions we can just treat $-L\frac{dI}{dt}$ as a voltage drop and it works fine. Most of the time thinking this way does not cause much of a problem. But technically it is not right!

We should consider where our magnetic flux came from. The magnetic flux was created by the current. It is a self-field. The current can't make a magnetic flux that would then modify that current. This self-flux won't make an electric field in the wire. So there is no electric field in the big loop, so there is no potential drop in that part of the circuit. It is just that $\oint \mathbf{E} \cdot d\mathbf{s} \neq 0$ because our field is not conservative. We had to take some energy to create the magnetic field.

Now, if you are doing simple circuit design, you can pretend you don't know about Faraday's law and this complication and just treat $-L\frac{dI}{dt}$ as though it were a voltage drop. But really it is just that going around the loop we should expect

$$\oint \mathbf{E} \cdot d\mathbf{s} = L\frac{dI}{dt}$$

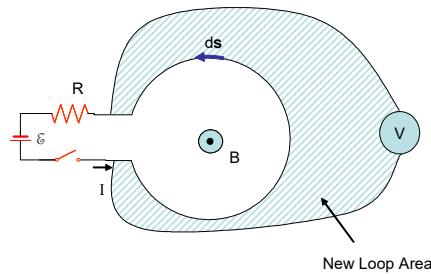
not

$$\oint \mathbf{E} \cdot d\mathbf{s} = 0$$

The danger is that if you are designing a complicated device that depends on there being an electric field in the inductor, your device will not work. We have *no external magnetic field*, our only magnetic field is the *self-field* which will not produce an electric field (or at least will form a very small electric field compared to the electric fields in the resistor and the battery, due to the small resistance in the real wire we use to make the big loop).

This is very subtle, and I struggle to remember this! Fortunately in most circuit design it does not matter. We just treat the inductor as though it were a true voltage drop.

I can make it even more exasperating by asking what you will see if you place a voltmeter across the inductor. What I measure is a “voltage drop” of LdI/dt , so maybe there is a voltage drop after all! But no, that is not right. The problem is that in introducing the voltmeter, we have created a new loop. For this loop, the field from our big loop *is* an external field. .



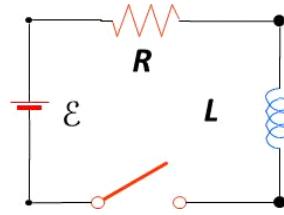
So the changing magnetic field through this voltmeter loop will produce an emf that will just match LdI/dt . And there will be an electric field—but it will be in the internal resistor in the voltmeter. And that is what you will measure!

Pick it up here

This may all seem very far fetched. But if you are designing radio communications you *want* to have a loss into the magnetic field, because that energy transferred to the magnetic field becomes your radio signal. This could be important!

The bottom line is that for non-conservative fields you need to be careful. If you are just designing simple circuits, you can just treat LdI/dt as though it were a voltage drop, but you may be badly burned by this if your system is more complicated, depending on the existence of a real electric field. You can see that if you are designing complicated sensing devices, you may need to deeply understand the underlying physics to get them to work. When in doubt, consult with a really good electrical engineer!

RL Circuits: Solving for the current as a function of time



The equation we found from Faraday's law or incorrectly from Kirchhoff's rule is

$$\mathcal{E} - IR - L \frac{dI}{dt} = 0 \quad (46.20)$$

This is a differential equation. We can solve it for the current. To do so, let's define a variable

$$x = \frac{\mathcal{E}}{R} - I$$

and then we see that

$$dx = -dI$$

Then we can write our differential equation as

$$\begin{aligned} \frac{\mathcal{E}}{R} - I - \frac{L}{R} \frac{dI}{dt} &= 0 \\ x + \frac{L}{R} \frac{dx}{dt} &= 0 \end{aligned}$$

and so

$$x = -\frac{L}{R} \frac{dx}{dt}$$

You might be able to guess the solution at this point from your M316 experience. But let's work it out as a review. We see that our x equation separates into

$$\frac{dx}{x} = -\frac{R}{L} dt$$

Integration yields

$$\begin{aligned} \int_{x_o}^x \frac{dx}{x} &= - \int_0^t \frac{R}{L} dt \\ \ln\left(\frac{x}{x_o}\right) &= -\frac{R}{L} t \end{aligned}$$

exponentiating both sides gives

$$\left(\frac{x}{x_o}\right) = e^{-\frac{R}{L} t}$$

Now we replace x with $\frac{\mathcal{E}}{R} - I$

$$\left(\frac{\frac{\mathcal{E}}{R} - I}{x_o}\right) = e^{-\frac{R}{L} t}$$

And because at $t = 0, I = 0$

$$\left(\frac{\mathcal{E}}{R} - I \right) = e^{-\frac{R}{L}t}$$

rearranging gives

$$I = \frac{\mathcal{E}}{R} \left(1 - e^{-\frac{Rt}{L}} \right) \quad (46.21)$$

or, defining another time constant

$$\tau = \frac{L}{R} \quad (46.22)$$

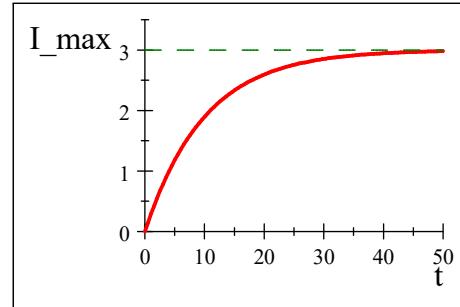
we have

$$I = \frac{\mathcal{E}}{R} \left(1 - e^{-\frac{t}{\tau}} \right) \quad (46.23)$$

We can see that

$$\frac{\mathcal{E}}{R} = I_{\max} \quad (46.24)$$

comes from Ohm's law. So just like with our capacitor-resistor circuit, we have a current that grows in time, approaching the maximum value we get after a time t which is much longer than τ .



You might expect that, like for a capacitor, there is an equation for an inductor who has a maximum current flowing but for which the current source is shorted (disconnected, and replaced with a resistanceless wire). The equation is

$$I = I_o e^{-\frac{t}{\tau}} \quad (46.25)$$

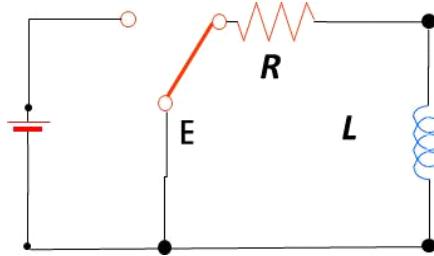
Magnetic Field Energy in Circuits

We found that just like with a RC circuit, we should expect there to be energy stored in

a *RL* circuit.

$$U_L = \frac{1}{2}LI^2 = \frac{1}{2}C(\Delta V)^2$$

Consider once again the *RL* circuit shown below.



Recall that the current in the right-hand loop decays exponentially with time according to the expression

$$I = I_o e^{-\frac{t}{\tau}}$$

where $I_o = E/R$ is the initial current in the circuit and $\tau = L/R$ is the time constant.

As an example problem, let's show that all the energy initially stored in the magnetic field of the inductor appears as internal energy in the resistor as the current decays to zero.

Recall that energy is delivered to the resistor

$$\frac{dU}{dt} = P = I^2 R$$

where I is the instantaneous current.

$$\begin{aligned}\frac{dU}{dt} &= I^2 R \\ \frac{dU}{dt} &= (I_o e^{-\frac{t}{\tau}})^2 R \\ \frac{dU}{dt} &= I_o^2 e^{-2\frac{t}{\tau}} R\end{aligned}$$

To find the total energy delivered to the resistor we integrate

$$\begin{aligned}dU &= I_o^2 e^{-2\frac{t}{\tau}} R dt \\ \int dU &= \int_0^\infty I_o^2 e^{-2\frac{t}{\tau}} R dt \\ U &= \int_0^\infty I_o^2 e^{-2\frac{t}{\tau}} R dt \\ U &= I_o^2 R \int_0^\infty e^{-2\frac{t}{\tau}} dt\end{aligned}$$

Use your calculator, or an integral table, or Maple, or your very good memory to recall that

$$\int e^{-ax} dx = -\frac{1}{a} e^{-ax}$$

If we let

$$a = -\frac{2}{\tau}$$

then we can obtain

$$U = -\frac{L}{2R} I_o^2 R e^{-2\frac{t}{\tau}} \Big|_0^\infty$$

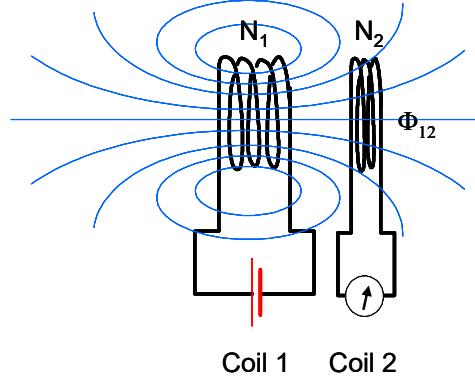
$$U = \frac{-L}{2} I_o^2 (0 - 1)$$

$$U = \frac{1}{2} I_o^2 L \quad (46.26)$$

which is the initial energy stored in the magnetic field. All of the energy that started in the inductor was delivered to the resistor.

Mutual Induction

Suppose we have two coils near each other. If either of the coils carries a current, will there be an induced current in the other coil?



We define Φ_{12} as the flux through coil 2 due to the current in coil 1. Likewise if the battery is placed on coil 2 we would have Φ_{21} , the flux through coil 1 due to the current in coil 2.

We define the mutual inductance

$$M_{12} = \frac{N_2 \Phi_{12}}{I_1} \quad (46.27)$$

BE CAREFUL! Not all books write the subscripts in the same order!

We can write the flux as

$$\Phi_{12} = \frac{M_{12} I_1}{N_2}$$

Then, using Faraday's law, we find the induced emf in coil 2

$$\begin{aligned}\mathcal{E}_2 &= -N_2 \frac{d\Phi_B}{dt} \\ &= -N_2 \frac{d}{dt} \left(\frac{M_{12} I_1}{N_2} \right) \\ &= -M_{12} \frac{d}{dt} (I_1)\end{aligned}$$

We state without proof the $M_{12} = M_{21}$. Then

$$\mathcal{E}_2 = -M \frac{dI_1}{dt}$$

Example : “Wireless” battery charger



Rechargeable Toothbrush with an inductive charger (Public Domain Image courtesy Jonas Bergsten)

A rechargeable toothbrush needs a connection that is not affected by water. We can use induction to form this connection. We need two coils. One coil is the base, the other the handle. The base carries current I . The base has length l and area A and N_B turns. The handle has N_H turns and completely covers the base solenoid. What is the mutual inductance?

Solution:

The magnetic field in the base solenoid is

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mathbf{B} \cdot \ell = \mu_o N_B I$$

or

$$B = \frac{\mu_o N_B I_B}{\ell}$$

Because the handle surrounds the base, the flux through the handle is the interior field

of the base. The flux is

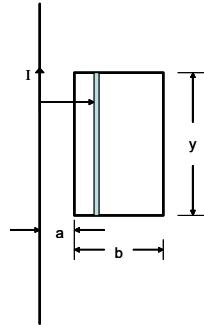
$$\Phi_{BH} = BA$$

The mutual inductance is

$$\begin{aligned} M &= \frac{N_H \Phi_{BH}}{I_B} \\ &= \frac{N_H B A}{I_B} \\ &= \frac{N_H \left(\frac{\mu_o N_B I_B}{\ell} \right) A}{I_B} \\ &= \mu_o \frac{N_H N_B A}{\ell} \end{aligned}$$

Example: Rectangular Loop and a coil

A rectangular loop of N close-packed turns is positioned near a long straight wire.



What is the coefficient of mutual inductance M for the loop-wire combination?

The basic equations are

$$M_{12} = \frac{N_2 \Phi_{12}}{I_1}$$

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_o I$$

$$\oint \mathbf{B} \cdot d\mathbf{A} = \Phi_B$$

The field from the wire

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_o I$$

Take the path to be a circle surrounding the wire then \mathbf{B} is constant along the path and

the direction of \mathbf{B} is tangent to the path.

$$\begin{aligned} B \oint ds &= \mu_o I \\ B 2\pi r &= \mu_o I \end{aligned}$$

or

$$B = \frac{\mu_o I}{2\pi r}$$

The flux through the rectangular loop is then perpendicular to the plane of the loop

$$\begin{aligned} \oint \mathbf{B} \cdot d\mathbf{A} &= \Phi_B \\ \Phi_B &= \int B y dr \\ &= \int_a^{b+a} \frac{\mu_o I}{2\pi r} y dr \\ &= \frac{\mu_o I y}{2\pi} \ln \frac{b+a}{a} \end{aligned}$$

then

$$M = N \frac{\mu_o y}{2\pi} \ln \frac{b+a}{a}$$

Suppose the loop has $N = 100$ turns, $a = 1\text{ cm}$, $b = 8\text{ cm}$, $y = 30\text{ cm}$, $\mu_o = 4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}$ what is the value of the mutual inductance?

$$M = N \frac{\mu_o y}{2\pi} \ln \frac{b+a}{a} = \frac{1.3183 \times 10^{-3}}{\text{A}} \text{ T m cm} = \frac{1.3183 \times 10^{-5}}{\text{A}^2} \frac{\text{m}^2}{\text{s}^2} \text{ kg}$$

$$H = \frac{1}{A^2} \frac{m^2}{s^2} \text{ kg}$$

Basic Equations

47 The Electromagnetic field

We started off our study of electricity and magnetism saying we would consider the environment made by a charge and how that environment affected a mover charge. Then we found that moving charges are affected by the environment created by other moving charges (currents). It is time to consider the overall environment created by both electric and magnetic fields acting together.

Fundamental Concepts

- The electric and magnetic fields are really different manifestations of the electromagnetic field. Which is manifest depends on our relative motion.
- The Galilean field transformations are

$$\vec{E}' = \vec{E}_{\text{charges}} + \vec{V}_{S'S} \times \vec{B}_{\text{environment}}$$

$$\vec{B}' = \vec{B}_{\text{magnet}} - \frac{1}{c^2} (\vec{V}_{S'S} \times \vec{E}_{\text{environment}})$$

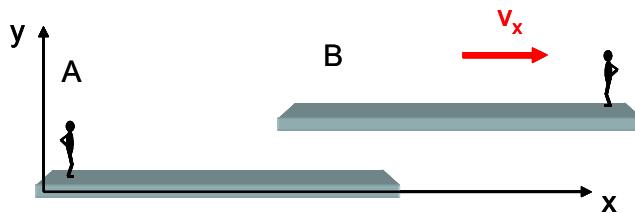
$$\vec{E} = \vec{E}'_{\text{charges}} - \vec{V}_{S'S} \times \vec{B}'_{\text{environment}}$$

$$\vec{B} = \vec{B}'_{\text{magnet}} + \frac{1}{c^2} (\vec{V}_{S'S} \times \vec{E}'_{\text{environment}})$$

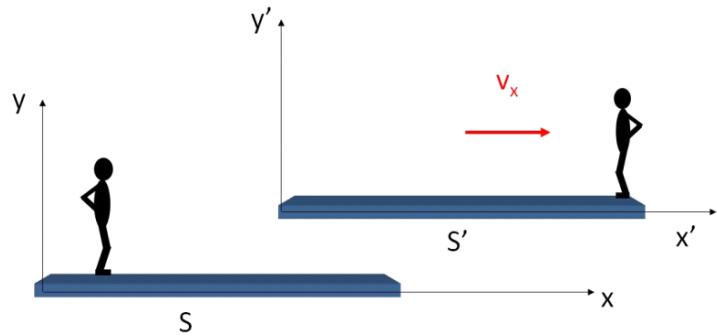
- Gauss' law for magnetic fields is $\oint \mathbf{B} \times d\mathbf{A} = 0$

Relative motion and field theory

Long ago in your study of physics we talked about relative motion when we discussed moving objects and Doppler shift. We considered two reference frames with a relative velocity v_i . We called them frame A and frame B



We need to return to relative motion, considering what happens when there are fields and charged particles involved. We will need to relabel our diagram to avoid confusion because now B will represent a magnetic field. So let's call the two reference frames S and S' . We will label each axis with a prime in the S' frame.



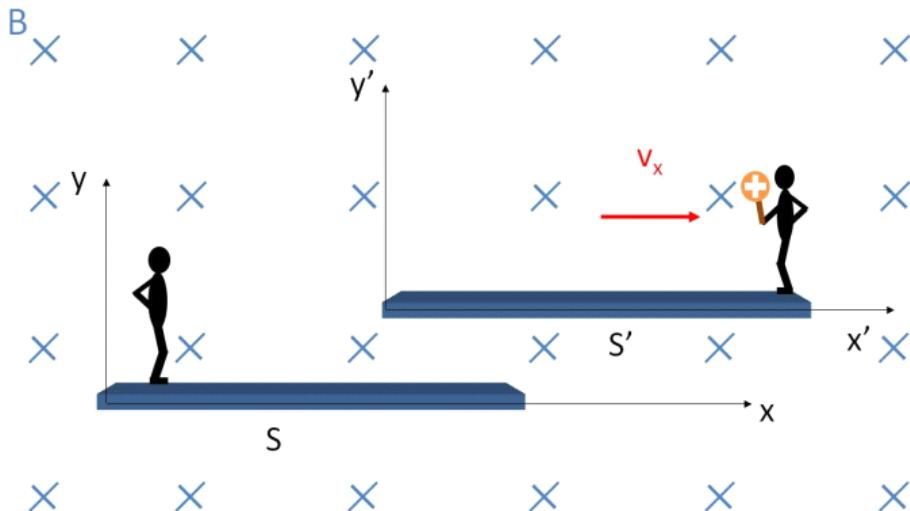
Question 223.47.1

Question
223.47.1.5

Question 223.47.2

Question 223.47.3

Now let's assume we have a magnetic field in the region of space where our two reference frames exist. Let's say that the magnetic field is stationary in frame S . This will be our environment. Let's also give a charge to the person in frame S' . This will be our mover charge.



Is there a force on the charge?

If we are with the person in reference frame S , then we must say yes. The charge is moving along with frame S' with a velocity $\vec{v} = V\hat{i}$ so there will be a force

$$\begin{aligned}\vec{F} &= q\vec{v} \times \vec{B} \\ &= qV\hat{i} \times B(-\hat{k}) \\ &= qVB\hat{j}\end{aligned}$$

in the \hat{j} direction.

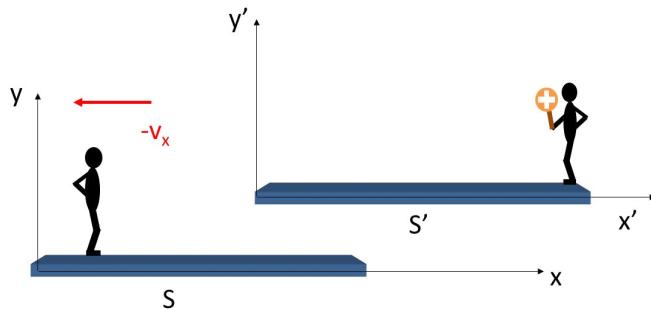
Now let's ride along with the person in frame S' . From this frame, the charge looks stationary. So $v = 0$ and

$$F = q(\mathbf{0}) \times \vec{B} = 0$$

Both can't be true! So which is it? Is there a force on the charge or not? Consider that the existence of a force is something we can test. A force causes motion to change in ways we can detect. (the person in frame S' would *feel* the pull on the charge he is holding). So ultimately we can perform the experiment and see that there really is a force. But where does the force come from?

Let's consider our fields. We have come to see fields as the source of electric and magnetic forces. Electric forces come from electric fields which come from environmental charges. Magnetic forces come from environmental magnetic fields which come from moving charges.

And here is the difficulty, we are having trouble recognizing when the charge is moving. We know from our consideration of relative motion that we could view this situation as frame S' moving to the right with frame S stationary, or frame S moving to the left with frame S' stationary. There is no way to say that only one of these views is correct. Both are equally valid.

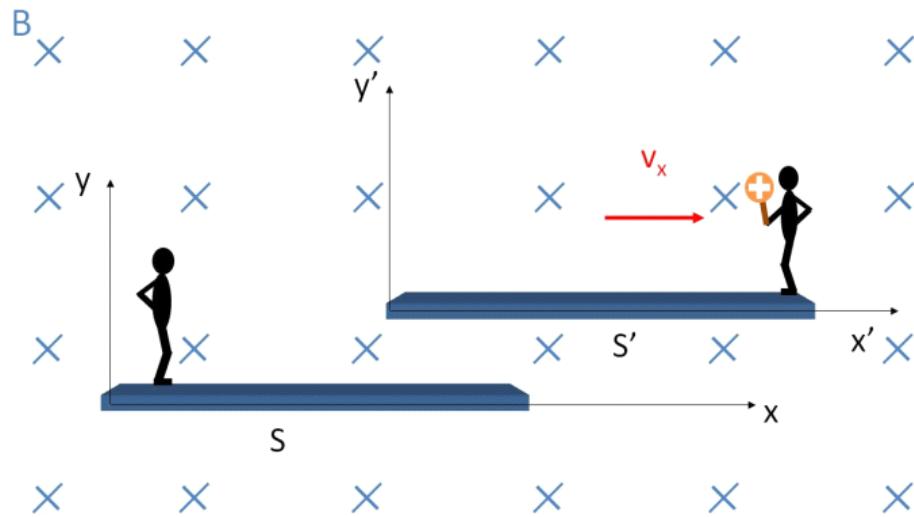


In our case, we are considering that person S sees a moving charge. We have learned

that moving charge will make *both* an electric field *and* a magnetic field. This is the situation from frame S . But person S' sees a static charge. This charge will *only* make an electric field. We need a way to resolve this apparent contradiction.

Galilean transformation

To resolve this difficulty, let's go back to forces. Here is our case of a constant magnetic field that is stationary in frame S with a charge in frame S' again.



We can't see fields, but we can see acceleration of a particle. Since by Newton's second law

$$F = ma$$

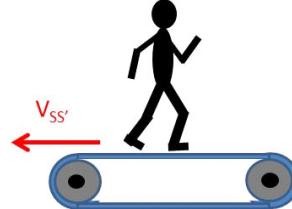
we will know if there is an acceleration and therefore we will know if there is a force! So are the forces and accelerations of a charged particle the same in each frame? Let's find out.

Remember from Dynamics or PH121 that the speed of a particle transforms like this

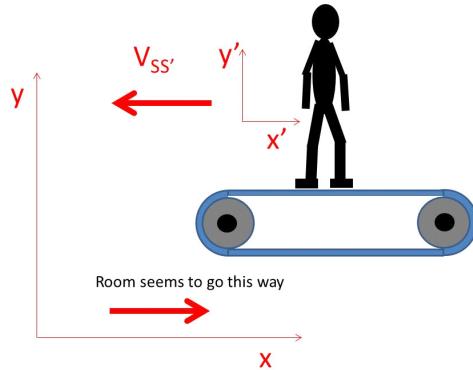
$$\begin{aligned}\vec{v}' &= \vec{v} - \vec{V}_{S'S} \\ \vec{v} &= \vec{v}' + \vec{V}_{S'S}\end{aligned}\tag{47.1}$$

where $V_{S'S}$ is the relative speed between the two frames. What this means is that if we have a particle moving with speed v' in frame S' and we observe this particle in frame S the speed of that particle will seem to be $\vec{v} = \vec{v}' + \vec{V}_{S'S}$. In our case, $\vec{V}_{SS'} = V_x \hat{i}$ so $\vec{v} = \vec{v}' + V_x \hat{i}$.

A quick example might help. Suppose we have a person in the gym running on a treadmill.

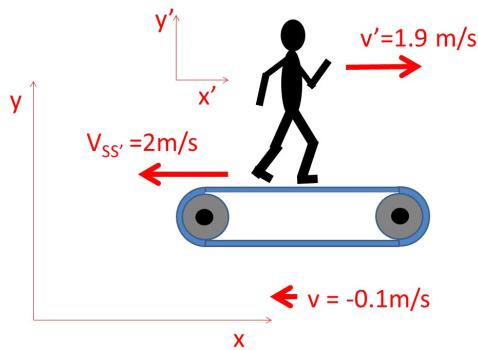


The treadmill track belt has a relative speed $\vec{V}_{S'S} = -2 \frac{\text{m}}{\text{s}} \hat{i}$ with respect to the room. We will say that the room is frame S . Then if we envision a reference frame riding along the treadmill, that would be frame S' . A person standing on the treadmill in frame S' sees themselves as not moving, and the rest of the room as moving the opposite direction.



The notation $V_{S'S}$ means the speed of the reference frame S' with respect to frame S or in our case the speed of the treadmill with respect to the room $\vec{V}_{S'S} = -2 \frac{\text{m}}{\text{s}} \hat{i}$.

Now suppose the person is running at speed $\vec{v}' = 1.9 \frac{\text{m}}{\text{s}} \hat{i}'$ on the tread mill in the tread mill frame S' .



What is his/her speed with respect to the room? It seems obvious that we take the two

speeds and add them.

$$\vec{v} = 1.9 \frac{\text{m}}{\text{s}} \hat{i}' - 2 \frac{\text{m}}{\text{s}} \hat{i} = -0.1 \frac{\text{m}}{\text{s}} \hat{i}$$

since the i and i' directions are the same.

The person is going to fall off the end of the treadmill unless they pick up the pace!

This example just used the second equation in our transformation.

$$\vec{v} = \vec{v}' + \vec{V}_{S'S}$$

likewise, if we want to know how fast the person is walking with respect to the treadmill frame, we take the room speed $\vec{v} = -0.1 \frac{\text{m}}{\text{s}} \hat{i}$ and subtract from it the treadmill/room relative speed $\vec{V}_{S'S} = -2 \frac{\text{m}}{\text{s}} \hat{i}$ to obtain

$$\vec{v}' = -0.1 \frac{\text{m}}{\text{s}} \hat{i} - \left(-2 \frac{\text{m}}{\text{s}} \hat{i} \right) = 1.9 \frac{\text{m}}{\text{s}} \hat{i} = 1.9 \frac{\text{m}}{\text{s}} \hat{i}'$$

Armed with the Galilean transform, we can find the acceleration by taking a derivative

$$\begin{aligned} \frac{d\vec{v}'}{dt} &= \frac{d\vec{v}}{dt} - \frac{d\vec{V}_{S'S}}{dt} \\ \frac{d\vec{v}}{dt} &= \frac{d\vec{v}'}{dt} + \frac{d\vec{V}_{S'S}}{dt} \end{aligned}$$

then

$$\begin{aligned} \vec{a}' &= \vec{a} - \frac{d\vec{V}_{S'S}}{dt} \\ \vec{a} &= \vec{a}' + \frac{d\vec{V}_{S'S}}{dt} \end{aligned}$$

but we will only consider constant relative motion³⁶, so

$$\frac{d\vec{V}_{S'S}}{dt} = 0$$

then both equations tell us

$$\vec{a}' = \vec{a}$$

This was a lot of work, but the end of all this talk about reference frames shows us that there *must be a force*

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

in both frame S and S' . The mass is the same in both frames, and so is the acceleration.

We can gain some insight into finding the mysterious missing force in frame S' by considering the net force in the case of both an electric and a magnetic field

$$\vec{F}_{net} = q \vec{E} + q \vec{v} \times \vec{B}$$

This was first written by Lorentz, so it is called the *Lorentz force*, and is usually written

³⁶ Accelerating reference frames are treated by General Relativity and are treated with the notation of contravariant and covariant vectors, which are beyond this course. They are taken up in a graduate level electricity and magnetism course.

as

$$\vec{\mathbf{F}}_{net} = q \left(\vec{\mathbf{E}} + \vec{\mathbf{v}} \times \vec{\mathbf{B}} \right)$$

Using this, let's consider the view point of each frame.

Going back to our two guys on different frames, In frame S , the person sees

$$\begin{aligned}\vec{\mathbf{F}} &= q \left(0 + \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}} \right) = qV_x \hat{i} \times B \left(-\hat{k} \right) \\ &= qVB \hat{j}\end{aligned}$$

and in frame S' the person sees

$$\vec{\mathbf{F}}' = q \left(\vec{\mathbf{E}}' + 0 \times \vec{\mathbf{B}}' \right) = q\vec{\mathbf{E}}'$$

It seems that the only way that $\vec{\mathbf{F}} = \vec{\mathbf{F}}'$ is that $\vec{\mathbf{E}}' \neq 0$ in the primed frame! So in frame S' our person must conclude that there is an external electric field that produces the force $\vec{\mathbf{F}}'$. In frame S the person is convinced that the magnetic field, $\vec{\mathbf{B}}$, is making the force. In frame S' the person is convinced that the electric field $\vec{\mathbf{E}}'$ is making the force.

Question 223.47.4

We can find the strength of this electric field by setting the forces equal

$$\begin{aligned}\vec{\mathbf{F}} &= \vec{\mathbf{F}}' \\ q\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}} &= q\vec{\mathbf{E}}'\end{aligned}$$

so

$$\vec{\mathbf{E}}' = \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}$$

and the direction must be

$$\vec{\mathbf{E}}' = V_{S'S} B \hat{j}$$

Question 223.47.5

Question 223.47.6

Our interpretation of this result is mind-blowing. It seems that whether we see a magnetic field or an electric field causing the force depends on our reference frame! The implication is that the electric and magnetic fields are not really two different things. They are one field viewed from different reference frames!

Another way to say what we have found might be that moving magnetic fields show up as electric fields.

So far we have been talking about external fields only. The field $\vec{\mathbf{B}}$ in our case study is created by some outside agent. So the field $\vec{\mathbf{E}}'$ observed in frame S' is also an environmental field. But the charge, itself, creates a field. So the total electric field in frame S' is the environmental field $\vec{\mathbf{E}}'$ plus the field due to the charge, itself $\vec{\mathbf{E}}_{self}$, or

$$\begin{aligned}\vec{\mathbf{E}}'_{tot} &= \vec{\mathbf{E}}_{self} + \vec{\mathbf{E}}'_{environment} \\ &= \vec{\mathbf{E}}_{self} + \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}_{environment}\end{aligned}$$

which we usually just write as

$$\vec{E}' = \vec{E}_{\text{self}} + \vec{V}_{S'S} \times \vec{B}_{\text{environment}}$$

We would predict that if we had a charge that is stationary in frame S and we rode along with frame S' that we would see a field

$$\vec{E} = \vec{E}'_{\text{self}} - \vec{V}_{S'S} \times \vec{B}'_{\text{environment}}$$

Of course, \vec{E}'_{self} can't create a force on the charge, because it is a self-field. So we only need to be concerned with \vec{E}'_{self} if we have other charges that could move. We could actually have a group of charges riding along with frame S' . In that case we would have an additional field E'_{charges} . We could write this as

$$\vec{E} = E'_{\text{charges in } S'} - \vec{V}_{S'S} \times \vec{B}'_{\text{environment}}$$

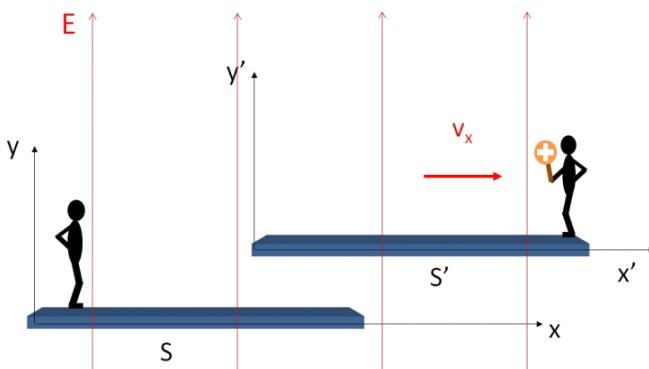
or just

$$\vec{E} = E'_{\text{charges}} - \vec{V}_{S'S} \times \vec{B}'_{\text{environment}}$$

What we have developed is important! We have an equation that let's us determine the electric field in a frame, given the fields measured in another frame.

We would expect that a similar thing would happen if we replaced the magnetic fields with electric fields. Suppose we have an electric field in the region of our frames and that this electric field is stationary with respect to frame S' this time. Will frame S see a magnetic field?

Question 223.47.7



To see that this is true, let's examine the case where we have no external fields, and we just have a charge moving along with frame S' . Then in frame S' we have the fields

$$\vec{E}' = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r}$$

$$\vec{B}' = 0$$

in frame S the electric field is

$$\begin{aligned}\vec{\mathbf{E}} &= \vec{\mathbf{E}}'_{\text{charges}} - \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}'_{\text{environment}} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{\mathbf{r}} + \vec{\mathbf{V}}_{S'S} \times \mathbf{0} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{\mathbf{r}}\end{aligned}$$

so

$$\vec{\mathbf{E}} = \vec{\mathbf{E}}' = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{\mathbf{r}}$$

We see the same electric field due to the point charge being there in both frames.

But in frame S we are expecting the person to see a magnetic field because to person S the charge is moving. Using the Biot-Savart law

$$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \frac{q \vec{\mathbf{v}} \times \hat{\mathbf{r}}}{r^2}$$

since our charge is moving along with the S' frame $\vec{\mathbf{v}} = \vec{\mathbf{V}}_{S'S}$ so

$$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \frac{q}{r^2} (\vec{\mathbf{V}}_{S'S} \times \hat{\mathbf{r}})$$

but we can rewrite this by rearranging terms

$$\begin{aligned}\vec{\mathbf{B}} &= \frac{\mu_0}{4\pi} \frac{q}{r^2} (\vec{\mathbf{V}}_{S'S} \times \hat{\mathbf{r}}) \\ &= (\vec{\mathbf{V}}_{S'S} \times \frac{\mu_0}{4\pi} \frac{q}{r^2} \hat{\mathbf{r}})\end{aligned}$$

which looks vaguely familiar. Let's multiply top and bottom by ϵ_0

$$\begin{aligned}\vec{\mathbf{B}} &= \left(\vec{\mathbf{V}}_{S'S} \times \frac{\mu_0 \epsilon_0}{4\pi} \frac{q}{r^2} \hat{\mathbf{r}} \right) \\ &= \left(\vec{\mathbf{V}}_{S'S} \times \mu_0 \epsilon_0 \left(\frac{1}{4\pi \epsilon_0} \frac{q}{r^2} \hat{\mathbf{r}} \right) \right) \\ &= \left(\vec{\mathbf{V}}_{S'S} \times \mu_0 \epsilon_0 (\vec{\mathbf{E}}') \right) \\ &= \mu_0 \epsilon_0 (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}')\end{aligned}$$

which is really quite astounding! Our B -fields have apparently always just been due to moving electric fields after all! Of course, we could have an additional magnet riding along with frame S' . To allow for that case, let's include a term $\vec{\mathbf{B}}'_{\text{magnet}}$.

$$\vec{\mathbf{B}}_{\text{total}} = \vec{\mathbf{B}}'_{\text{magnets in } S'} + \mu_0 \epsilon_0 (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}'_{\text{environment}})$$

or just

$$\vec{\mathbf{B}} = \vec{\mathbf{B}}'_{\text{magnets}} + \mu_0 \epsilon_0 (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}'_{\text{environment}})$$

and we would expect that if we worked this problem from the other frame's point of view we would likewise find

$$\vec{\mathbf{B}}' = \vec{\mathbf{B}}_{\text{magnet}} - \mu_0 \epsilon_0 (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}_{\text{environment}})$$

where the minus sign comes from the relative velocity being in the other direction.

Again \vec{B}_{magnet} is a self-field. It won't move the magnet creating it, but it might be important if we have a second magnet in our experiment. Then \vec{B}_{magnet} would cause a force on this second magnet.

Once again we have found a way to find a field, the magnetic field this time, in one frame if we know the fields on another frame! We call this sort of equation a *transformation*.

We should take a moment to look at the constants $\mu_0 \epsilon_0$. Let's put in their values

$$\begin{aligned}\mu_0 \epsilon_0 &= \left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2}\right) \left(4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}\right) \\ &= 1.1121 \times 10^{-17} \frac{\text{s}^2}{\text{m}^2}\end{aligned}$$

This is a very small number, and it may not appear to be interesting. We can see that the additional magnetic fields due to the movement of the charges can be quite small unless the electric field is large or the relative speed is large (or both). So much of the time this additional field due to the moving charge is negligible. But let's calculate

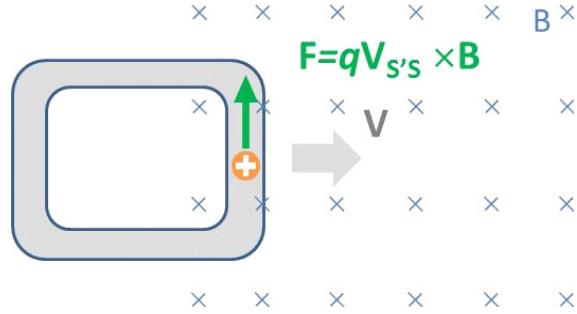
$$\begin{aligned}\frac{1}{\sqrt{\mu_0 \epsilon_0}} &= \frac{1}{\sqrt{\left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2}\right) \left(4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}\right)}} \\ &= 2.9986 \times 10^8 \frac{\text{m}}{\text{s}} \\ &= c\end{aligned}$$

This is the speed of light! It even has units of m/s. This seems an amazing coincidence—too amazing. And this was one of the clues that Maxwell used to discover that light is a wave in what we will now call the *electromagnetic field* (because they are different aspects of one thing).

We can write the transformation equations for the fields as

$$\begin{aligned}\vec{E}' &= \vec{E}_{\text{charges}} + \vec{V}_{S'S} \times \vec{B}_{\text{environment}} \\ \vec{B}' &= \vec{B}_{\text{magnet}} - \frac{1}{c^2} (\vec{V}_{S'S} \times \vec{E}_{\text{environment}}) \\ \vec{E} &= \vec{E}'_{\text{charges}} - \vec{V}_{S'S} \times \vec{B}'_{\text{environment}} \\ \vec{B} &= \vec{B}'_{\text{magnet}} + \frac{1}{c^2} (\vec{V}_{S'S} \times \vec{E}'_{\text{environment}})\end{aligned}$$

Let's do a problem. Suppose we have a metal loop moving into an area where there is a magnetic field as shown. Let's show that there is a force on charges in this loop no matter what frame we consider. First, let's consider the frame where the magnetic field is stationary and the loop moves.



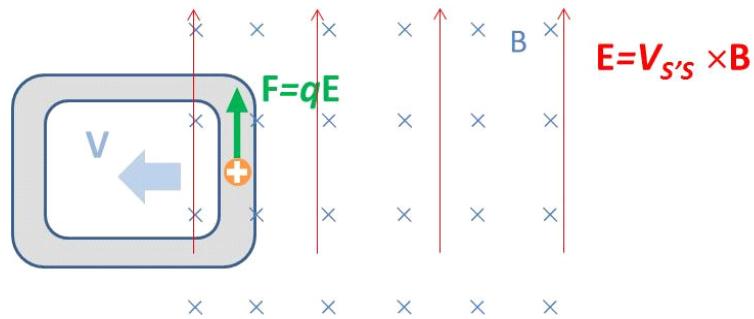
There should be an upward force on the positive charge because the charge is moving in a magnetic field. Let's say that "up" is the \hat{j} direction and that "to the right" is the \hat{i} direction. Then The Lorentz force is

$$\begin{aligned}\vec{F} &= q(\vec{E} + \vec{v} \times \vec{B}) \\ &= q(\vec{E} + \vec{V}_{S'S} \times \vec{B})\end{aligned}$$

Now $\vec{V}_{S'S}$ means the speed of the reference frame S' with respect to frame S . That is $+V\hat{i}$. And there is no electric field in frame S , so

$$\begin{aligned}\vec{F} &= q(\vec{E} + \vec{V}_{S'S} \times \vec{B}) \\ &= q(0 + V\hat{i} \times B(-\hat{k})) \\ &= q(V\hat{i} \times B(-\hat{k})) \\ &= qVB\hat{j}\end{aligned}$$

Now suppose we change reference frames so we are riding along with the loop in frame, S' . In this frame, the loop is not moving, and the magnetic field is moving by us the opposite direction. We'll call this the "prime frame." We should get the same force if we change frames to ride along with the loop.



Let's use our transformations to find the E and B -fields in the new reference frame.

Then

$$\begin{aligned}\vec{\mathbf{E}}' &= \vec{\mathbf{E}}_{\text{self-charge}} + \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}_{\text{environment}} \\ \vec{\mathbf{B}}' &= \vec{\mathbf{B}}_{\text{magnet}} - \frac{1}{c^2} (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}_{\text{environment}})\end{aligned}$$

so in the prime frame we have an electric field

$$\vec{\mathbf{E}}' = \vec{\mathbf{E}}_{\text{charges}} + \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}_{\text{environment}}$$

and in particular, we have an external field

$$\vec{\mathbf{E}}'_{\text{environment}} = \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}_{\text{environment}}$$

(we left off the $\vec{\mathbf{E}}_{\text{charge}}$ because it can't move the charge that made it, so it is not part of the force).

Note that $\vec{\mathbf{V}}_{S'S}$ is the speed of the primed frame as viewed from the unprimed frame.

So $\vec{\mathbf{V}}_{S'S} = +V\hat{i}$

$$\begin{aligned}\vec{\mathbf{E}}' &= V(\hat{i}) \times B(-\hat{k}) \\ &= VB\hat{j}\end{aligned}$$

That is our electric field in the primed frame.

The magnetic field in the primed frame is given by

$$\vec{\mathbf{B}}' = \vec{\mathbf{B}}_{\text{magnet}} - \frac{1}{c^2} (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}_{\text{environment}})$$

but there is no external electric field in the unprimed frame, so

$$\begin{aligned}\vec{\mathbf{B}}' &= \vec{\mathbf{B}}_{\text{magnet}} - \frac{1}{c^2} (\vec{\mathbf{V}}_{S'S} \times 0) \\ &= \vec{\mathbf{B}}_{\text{magnet}}\end{aligned}$$

where here “magnet” means what ever is making the magnetic field in the unprimed frame. Something must be there making the field, and it is not our charge. It could be an electromagnet, or a permanent magnet, we have not been told. But it is not our charge, so we know $\vec{\mathbf{B}}_{\text{magnet}}$ must be there and can act on our charge. So

$$\vec{\mathbf{B}}' = \vec{\mathbf{B}}$$

The magnetic field in the primed frame is just the same as the magnetic field we see in the unprimed frame. Then in the primed frame the Lorentz force is

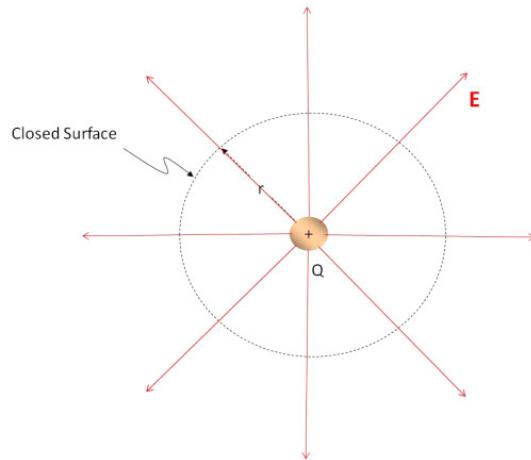
$$\begin{aligned}\vec{\mathbf{F}}' &= q(\vec{\mathbf{E}}' + \vec{\mathbf{v}} \times \vec{\mathbf{B}}') \\ &= q(VB\hat{j} + \mathbf{0} \times \vec{\mathbf{B}}) \\ &= qVB\hat{j}\end{aligned}$$

Which is exactly the same force (magnitude and direction) as we got in the unprimed frame.

Field Laws

A “law” in physics is a mathematical statement of a physical principal or theory. We have been collecting laws for what we will now call the *electromagnetic field theory*. Let’s review:

Gauss’ law



We found that the electric flux through an imaginary closed surface that incloses some charge is

$$\Phi_E = \oint \mathbf{E} \cdot d\mathbf{A} = \frac{Q_{in}}{\epsilon_0}$$

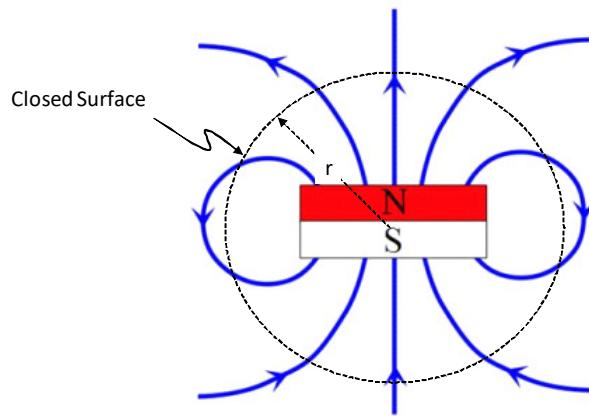
We called this Gauss’ law.

Question 223.47.8

But consider the situation with a magnet. We can define a magnetic flux just like we defined the electric flux. And now we know they must be related. Is there a Gauss’ law for magnetism? Let’s consider the magnetic flux.

$$\Phi_B = \oint \mathbf{B} \cdot d\mathbf{A}$$

This should be proportional to the number of “magnetic charges” inclosed in the surface.

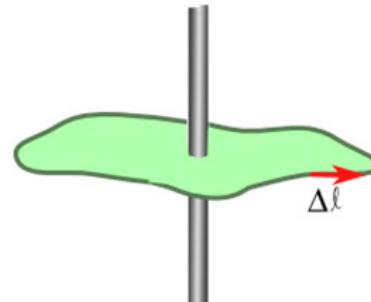


We can see that every field line that leaves comes back in. That is how we defined zero net flux, so

$$\Phi_B = \oint \mathbf{B} \cdot d\mathbf{A} = 0$$

Which would tell us that there are no free “magnetic charges” or no single magnetic poles. A single magnetic pole is called a *monopole* and indeed we have never discovered one. These two forms of Gauss’ law form the first two of our electromagnetic field equations.

The differences between them have to do with the fact that magnetic fields are due to moving charges.



We have a third electromagnetic field law, Ampere’s law. We found Ampere’s law by integrating around a closed loop with a current penetrating the loop.

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_o I_{\text{through}}$$

We also know Faraday's law

$$\mathcal{E} = \oint \mathbf{E} \cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$

which told us that changing magnetic fields created an electric field. We have found that the opposite must be true, that a changing electric field must create a magnetic field. We express this as

$$\oint \mathbf{B} \cdot d\mathbf{s} \propto \frac{d\Phi_E}{dt}$$

Which gives two expressions for $\oint \mathbf{B} \cdot d\mathbf{s}$. But we have yet to show that this equation is true. That is the subject of our next lecture. If we can accomplish this, we will have a complete set of field equations that describe how the electromagnetic field works. In the following lecture we will complete the set of field equations, and then in the next lecture we will show that we get electromagnetic waves from these equations.

Basic Equations

Rules for finding fields in different coordinate systems

$$\begin{aligned}\vec{\mathbf{E}}' &= \vec{\mathbf{E}}_{\text{charges}} + \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}_{\text{environment}} \\ \vec{\mathbf{B}}' &= \vec{\mathbf{B}}_{\text{magnet}} - \frac{1}{c^2} (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}_{\text{environment}}) \\ \vec{\mathbf{E}} &= \vec{\mathbf{E}}'_{\text{charges}} - \vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{B}}'_{\text{environment}} \\ \vec{\mathbf{B}} &= \vec{\mathbf{B}}'_{\text{magnet}} + \frac{1}{c^2} (\vec{\mathbf{V}}_{S'S} \times \vec{\mathbf{E}}'_{\text{environment}})\end{aligned}$$

Gauss' law for magnetic fields

$$\Phi_B = \oint \mathbf{B} \times d\mathbf{A} = 0$$

48 Field Equations and Waves in the Field

We started this class with a study of waves. We learned about optics, and finally electromagnetic field theory. In this lecture we will take on a case study that involves all three. We will have come full circle and in the process, hopefully understand all three topics a little better.

Fundamental Concepts

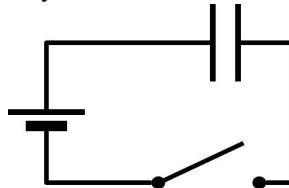
- Changing electric fields produce magnetic fields
- A changing electric flux is described as a displacement current $I_d = \varepsilon_o \frac{d\Phi_E}{dt}$
- The complete version of Ampere's law is $\oint \mathbf{B} \cdot d\ell = \mu_o (I + I_d)$
- Maxwell's equations give a complete classical picture of electromagnetic fields
- Maxwell's equations plus the Lorentz force describe all of electrodynamics.

Displacement Current

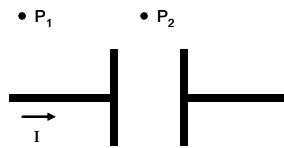
Last time we listed Ampere's law as one of the basic field equations. But we did not discuss it at all. That is because we were saving it for our discussion in this lecture. We need to look deeply into Ampere's law. Here is what we have for Ampere's law so far

$$\oint \mathbf{B} \cdot d\ell = \mu_o I_{\text{through}}$$

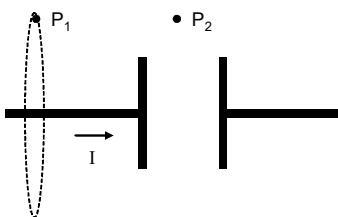
To see why we need to consider it further, let's do a hard problem with Ampere's law. Let's set up a circuit with a battery a switch and a circular plate capacitor in the wire.



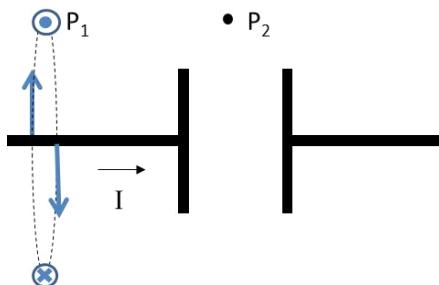
Using this circuit, let's calculate the magnetic field using Ampere's law. Here is a detailed diagram of the capacitor.



I could find the magnetic field using the Biot-Savart equation, but that would be hard. I don't know how to solve the resulting integral. So let's try Ampere's law. Let's start at P_1 . We add in an imaginary surface at P_1 . I will choose a simple circular surface.



We have done this before. If we choose P_1 so that it is far from the capacitor, then we know what the magnetic field will look like.



Right at P_1 it will be out of the page. We also know that for a long straight wire, the field magnitude does not change as we go around the wire, so we can write our integral as

$$\oint \mathbf{B} \cdot d\ell = B \oint d\ell = B2\pi r = \mu_o I$$

so

$$B2\pi r = \mu_o I$$

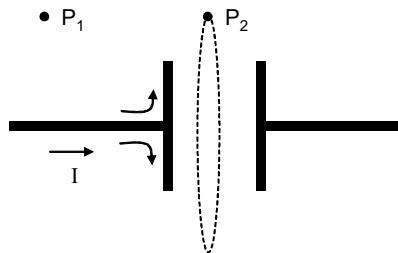
so the field is

$$B = \frac{\mu_o I}{2\pi r}$$

which is very familiar, just the equation for a field from a long straight wire.

Question 223.48.1

Now Let's try this at P_2 . What would we expect? Will the magnetic field change much as we pass by the capacitor?



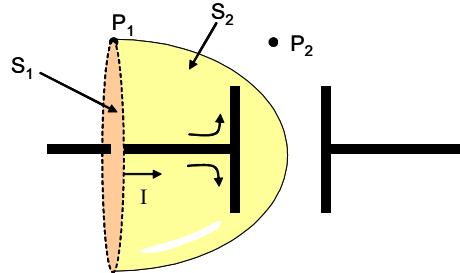
Again we could use Biot-Savart, but think about what the current does at the plate. It would be very hard to do the integration!. So again let's try Ampere's law. If we use the same size surface

$$\oint \mathbf{B} \cdot d\ell = B \oint d\ell = B2\pi r$$

but this is equal to $\mu_o I_{\text{through}}$. There is no I going through the capacitor! so

$$B2\pi r = 0 \quad (48.1)$$

and this would give $B = 0$. But, our wires are not really ideal and infinitely long. And even if they were, would we really expect the field to be zero if we just have a small gap in our capacitor? It get's even worse!



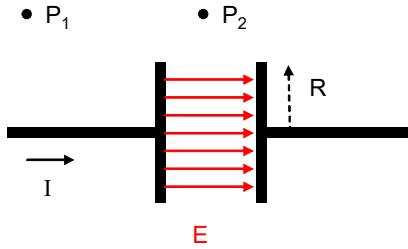
Ampere's law tells us we need a surface, but it does not say it has to be a circular surface. In fact, we could use the strange surface labeled S_2 in the figure above. This is a perfectly good surface to associate with the loop at P_1 . So this gives us

$$\oint \mathbf{B} \cdot d\ell = \mu_o I = 0$$

at P_1 ! So we have two different results with Ampere's law for the same point. This can't be!

Question 223.48.2

Ampere knew this was a problem, but did not find a solution. Maxwell solved this. He asked himself, what was different inside the capacitor that might be making a difference. Of course, there is an electric field inside the capacitor!



We know that in the limit that the plates can be considered to be very big the field is approximately

$$E = \frac{\eta}{\varepsilon_o} = \frac{Q}{\pi R^2 \varepsilon_o}$$

but we know that the charge is changing in time once the switch is thrown. We can find the rate of change of the field, then

$$\frac{dE}{dt} = \frac{1}{\pi R^2 \varepsilon_o} \frac{dQ}{dt}$$

By definition

$$I = \frac{dQ}{dt}$$

is a current, but what current? It must be the current that is supplying the charge to the capacitor. That current is what is changing the Q in the capacitor, and it is the Q separation that is making the field. So the time derivative of the electric field is

$$\frac{dE}{dt} = \frac{I}{\pi R^2 \varepsilon_o}$$

where I is the current in the wire, and only if the wire current is zero will there be no change in the electric field.

This gives us an idea. A changing electric field creates a magnetic field. Suppose this changing electric field created a magnetic field like the current does? It would as though there were a current with a value

$$I_d = \pi R^2 \varepsilon_o \frac{dE}{dt} \quad (48.2)$$

It doesn't really cause a current in the capacitor. What really happens is that the changing electric field is creating a magnetic field. But that magnetic field is just like the field that a current would create. So we can (somewhat incorrectly) say that the changing electric field has created something like a current in the capacitor. But no charge is crossing the capacitor.

Note that in this we have the area of the plate, $A_{plate} = \pi R^2$ multiplied by the time rate of change of the electric field. Also note, that in our approximation for our capacitor,

there is only an electric field inside the plates. So, remembering electric flux,

$$\Phi_E = \int \mathbf{E} \cdot d\mathbf{A}$$

our flux through the surface at P_2 would be

$$\begin{aligned}\Phi_E &= EA \\ &= \pi R^2 E\end{aligned}$$

so we can identify

$$\pi R^2 dE = A_{plate} dE = d\Phi_E$$

as a small amount of *electric* flux. Then our equivalent current will be

$$I_d = \varepsilon_o \frac{d\Phi_E}{dt} \quad (48.3)$$

Maxwell decided that, since this looked like equivalent to a current, he would call it a current and include it in Ampere's law.

$$\begin{aligned}\oint \mathbf{B} \cdot d\ell &= \mu_o (I + I_d) \\ &= \mu_o \left(I + \varepsilon_o \frac{d\Phi_E}{dt} \right)\end{aligned}$$

but remember it is not really a current. What we have is a changing electric field that is making a magnetic field *as though there were a current* I_d . We can try this on or capacitor problem. We have done our capacitor problem for S_1 where we expect $\frac{d\Phi_E}{dt} \approx 0$ so our original calculation stands

$$B_{S_1} = \frac{\mu_o I}{2\pi r}$$

but now we know that if we use S_2 we have $\frac{d\Phi_E}{dt} \neq 0$, and we realize that at P_2 the current $I = 0$ so

$$\oint \mathbf{B} \cdot d\ell = \mu_o \left(0 + \varepsilon_o \frac{d\Phi_E}{dt} \right)$$

and for our geometry we found $\frac{d\Phi_E}{dt}$

$$\oint \mathbf{B} \cdot d\ell = \mu_o \left(0 + \pi R^2 \varepsilon_o \frac{dE}{dt} \right)$$

and we calculated $\frac{dE}{dt}$ so we can substitute it in

$$\oint \mathbf{B} \cdot d\ell = \mu_o \left(0 + \pi R^2 \varepsilon_o \frac{I}{\pi R^2 \varepsilon_o} \right)$$

where we remember that the current I is the current making the electric field—the current in the wire. Then we have

$$B 2\pi r = \mu_o (0 + I)$$

and our field is

$$B = \frac{\mu_o I}{2\pi r}$$

which is just what we found using S_1 . Maxwell seems to have saved the day! There is no dip in the magnetic field magnitude.

Question 223.48.3

There is one more fix we will have to do to Ampere's law eventually. We found this form of Ampere's law with the capacitor empty—not even containing air. But we could do the same derivation with a dielectric filled capacitor. We also could have magnetic materials involved.

But what we have done so far is really a momentous result. We have shown that, indeed, we should have an equation that provides symmetry with Faraday's law. We suspected that

$$\oint \mathbf{B} \cdot d\mathbf{s} \propto \frac{d\Phi_E}{dt}$$

and we can write the constants of proportionality as

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_o \epsilon_o \frac{d\Phi_E}{dt}$$

but because we have $\oint \mathbf{B} \cdot d\mathbf{s}$ also in Ampere's law, we can combine the two to yield

$$\begin{aligned} \oint \mathbf{B} \cdot d\mathbf{s} &= \mu_o (I + I_d) \\ &= \mu_o \left(I + \epsilon_o \frac{d\Phi_E}{dt} \right) \end{aligned}$$

This is the last of our field equations. It is called the Maxwell-Ampere law.

Let's use this to solve for the magnetic field inside the capacitor. A changing electric field will make a magnetic field.

Take a surface inside the plates that is a circle of radius $r < R$. Then

$$\oint \mathbf{B} \cdot d\mathbf{s} = B 2\pi r$$

and from our modified Ampere's equation

$$\begin{aligned} \oint \mathbf{B} \cdot d\mathbf{s} &= \mu_o (I + I_d) \\ &= \mu_o \left(I + \epsilon_o \frac{d\Phi_E}{dt} \right) \end{aligned}$$

so

$$\begin{aligned} B 2\pi r &= \mu_o \left(0 + \epsilon_o \frac{d\Phi_E}{dt} \right) \\ &= \pi r^2 \mu_o \epsilon_o \frac{dE}{dt} \\ &= \pi r^2 \mu_o \epsilon_o \frac{I}{\pi R^2 \epsilon_o} \\ &= \mu_o \frac{r^2 I}{R^2} \end{aligned}$$

so

$$B = \mu_o \frac{rI}{2\pi R^2} \quad (48.4)$$

We should pause to realize what we have just done. We have shown that, indeed, a changing electric field can produce a magnetic field. This statement is a profound look at the way the universe works!

Maxwell Equations

We have developed a powerful set of understanding equations for electricity and magnetism. Maxwell summarized our knowledge in a series of four equations

$$\begin{aligned} \oint \mathbf{E} \cdot d\mathbf{A} &= \frac{Q_{in}}{\epsilon_0} && \text{Gauss's law for electric fields} \\ \oint \mathbf{B} \cdot d\mathbf{A} &= 0 && \text{Gauss's law for magnetic fields} \\ \oint \mathbf{E} \cdot d\mathbf{s} &= -\frac{d\Phi_B}{dt} && \text{Faraday's law} \\ \oint \mathbf{B} \cdot d\mathbf{s} &= \mu_0 I + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt} && \text{Ampere-Maxwell Law} \end{aligned} \quad (48.5)$$

If we have a dielectric, we might see these written as[?]

$$\begin{aligned} \oint \mathbf{E} \cdot d\mathbf{A} &= \frac{Q_{in}}{\epsilon_0 \kappa} && \text{Gauss's law for electric fields} \\ \oint \mathbf{B} \cdot d\mathbf{A} &= 0 && \text{Gauss's law for magnetic fields} \\ \oint \mathbf{E} \cdot d\mathbf{s} &= -\frac{d\Phi_B}{dt} && \text{Faraday's law} \\ \oint \mathbf{B} \cdot d\mathbf{s} &= \mu_0 \kappa_m (I + \epsilon_0 \kappa \frac{d\Phi_E}{dt}) && \text{Ampere-Maxwell Law} \end{aligned} \quad (48.6)$$

Since we have all had multivariate calculus, we may also see these written as

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0} && \text{Gauss's law for electric fields} \\ \nabla \cdot \mathbf{B} &= 0 && \text{Gauss's law for magnetic fields} \\ \nabla \times \mathbf{E} &= -\frac{d\mathbf{B}}{dt} && \text{Faraday's law} \\ c^2 \nabla \times \mathbf{B} &= \frac{\mathbf{J}}{\epsilon_0} + \frac{d\mathbf{E}}{dt} && \text{Ampere-Maxwell Law} \end{aligned} \quad (48.7)$$

I'll let you remember the process to do the translation from $\oint \mathbf{B} \cdot d\mathbf{A}$ to $\nabla \cdot \mathbf{B}$.

But we are familiar with all of these equations now. These four equations are the basis of all of classical electrodynamics. In an electromagnetic problem, we find the fields using the Maxwell equations to find the fields, and then apply the fields to find the Lorentz forces

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \quad (48.8)$$

It turns out that these four equations strongly imply that there can be waves in the fields. Maxwell took the hint that $\mu_0 \epsilon_0$ was related to c , the speed of light and he thought

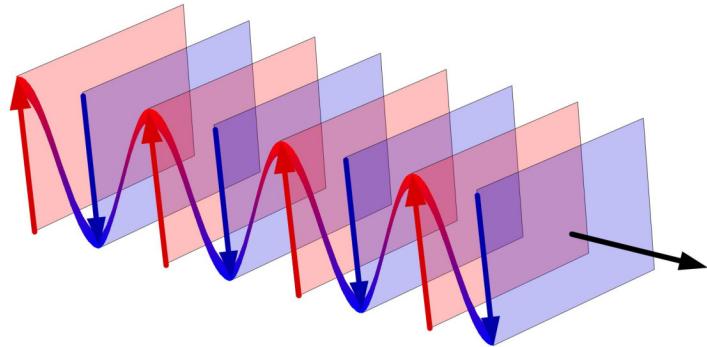
that light might be a wave in the electromagnetic field. We know about waves. We can describe a wave by looking for a surface of constant amplitude—a wave crest. We already know from our study of optics that these waves are what we call light. A point source will cause spherical surfaces of constant amplitude. A half-wave antenna makes a toroidal shaped wave front. We will not deal with spherical or worse wave shapes. Unfortunately, many antennas send out complicated wave patterns that take spherical harmonics to describe well. That is beyond the math we want to do in this course. We will stick to simple shapes. But we will see how waves in the electromagnetic field describe light in our next lecture.

49 Waves in the Field

We started this class with a study of waves. We learned about optics, and finally electromagnetic field theory. In this lecture we will take on a case study that involves all three. We will have come full circle and in the process, hopefully understand all three topics a little better.

Fundamental Concepts

- Maxwell's equation lead directly to the liner wave equation for both the electric and the magnetic field with the speed of light being the speed of the waves.
- The magnitude of the E and B fields are related in an electromagnetic wave by $E_{\max} = cB_{\max}$



A representation of a plane wave. Remember that the planes are really of infinite extent. Image is public domain.

Let's picture our wave front far from the source. No matter what the total shape, if we look at a small patch of the fields far away, they will look like the plane wave in the last figure. Since this is a useful and common situation (except if you use lasers), we will perform some calculations assuming plane wave geometry.

We will assume we are in empty space, so the charge q and current I will both be zero.

Then our Maxwell Equations become

$$\begin{aligned}
 \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} &= 0 && \text{Gauss's law for electric fields} \\
 \oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}} &= 0 && \text{Gauss's law for magnetic fields} \\
 \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} &= -\frac{d\Phi_B}{dt} && \text{Faraday's law} \\
 \oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} &= \epsilon_0 \mu_0 \frac{d\Phi_E}{dt} && \text{Ampere-Maxwell Law}
 \end{aligned} \tag{49.1}$$

Our goal is to show that these equations tell us that we can have waves in the field. To do this, we will show that Maxwell's equations really contain the linear wave equation within them. As a reminder, here is the linear wave equation

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2}$$

it is a second order differential equation where the left side derivatives are taken with respect to position, and the right side derivatives are taken with respect to time. The quantity, v , is the wave speed. In this form of the equation y is the displacement of a medium. Our medium will be the electromagnetic field.

Far Board

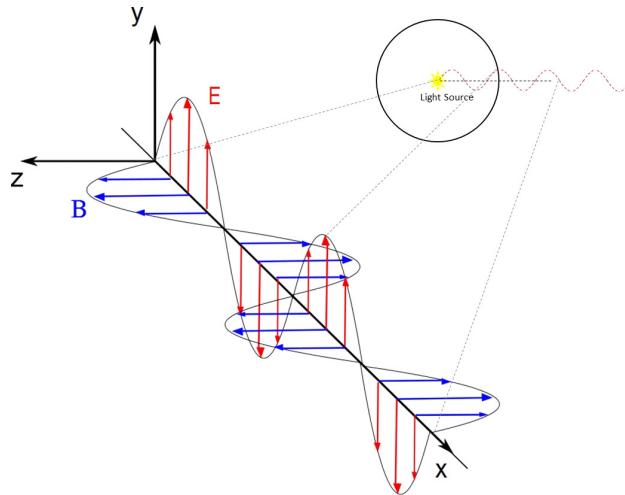
Rewriting of Faraday's law

Skip this

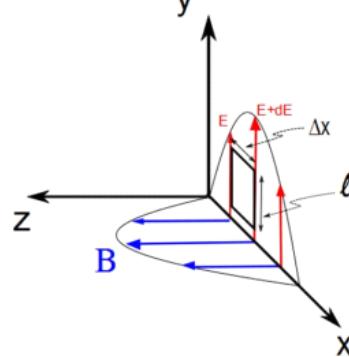
Let's start with Faraday's law

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d\Phi_B}{dt} \tag{49.2}$$

Given our geometry, we can say the wave is traveling in the x direction with the $\vec{\mathbf{E}}$ field positive in the y direction. From our discussion of displacement currents we have a strong hint that the $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ fields will be perpendicular. So let's take the magnetic field as positive in the z direction. So as the light wave moves from the source along a line we could draw the $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ fields something like this.



Let's take a small rectangle of area to find $\oint \vec{E} \cdot d\vec{s}$



The top and bottom of the rectangle don't contribute because $\vec{E} \cdot d\vec{s} = 0$ along these paths. On the sides, the field is either in the $d\vec{s}$ or it is in the opposite direction. So

$$\oint \vec{E} \cdot d\vec{s} = \oint E ds$$

or

$$\oint \vec{E} \cdot d\vec{s} = - \oint E ds$$

along the sides. Let's say we travel counter-clockwise along the loop. Then the left side will be negative and the right side will be positive.

$$\oint \vec{E} \cdot d\vec{s} = \int_{right} E ds - \int_{left} E ds$$

On the left side, we are at a position x away from the axis, and on the right side we are a position $x + \Delta x$ away from the axis. Then the field of the left side is $E(x, t)$ and the

field on the right hand side is approximately

$$E(x + \Delta x, t) \approx E(x, t) + \frac{\partial E}{\partial x} \Delta x \quad (49.3)$$

so if our loop is small, then ℓ is small and E won't change much so we can write approximately

$$\oint \vec{E} \cdot d\vec{s} = \int_{right} Eds - \int_{left} Eds \quad (49.4)$$

$$\approx E(x + \Delta x, t) \ell - E(x, t) \ell \quad (49.5)$$

$$= \left(E(x, t) + \frac{\partial E}{\partial x} \Delta x \right) \ell - E(x, t) \ell$$

$$= \left(E(x, t) + \frac{\partial E}{\partial x} \Delta x \right) \ell - E(x, t) \ell$$

$$= \ell \frac{\partial E}{\partial x} \Delta x \quad (49.6)$$

So far then, Faraday's law³⁷

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

becomes

$$\ell \frac{\partial E}{\partial x} \Delta x = -\frac{d\Phi_B}{dt}$$

Let's move on to the right hand side of Faraday's law. We need to find Φ_B so that we can find the time rate of change of the flux. We can say that B is nearly constant over such a small area, so

$$\begin{aligned} \Phi_B &= \mathbf{B} \cdot \mathbf{A} \\ &= BA \cos \theta \\ &= BA \\ &= B\ell \Delta x \end{aligned}$$

where here Δx means "a small distance" as it did above. Then

$$\begin{aligned} \frac{d\Phi_B}{dt} &= \frac{d}{dt} (B\ell \Delta x) \\ &= \ell \Delta x \frac{dB}{dt} \Big|_{x \text{ constant}} \\ &= \ell \Delta x \frac{\partial B}{\partial t} \end{aligned}$$

where we have held x constant because we are not changing our small area, so Faraday's law

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

³⁷ We need ds to be very small, much smaller than the wavelength of the wave.

becomes

$$\begin{aligned}\ell \frac{\partial E}{\partial x} \Delta x &= -\ell \Delta x \frac{\partial B}{\partial t} \\ \frac{\partial E}{\partial x} &= -\frac{\partial B}{\partial t}\end{aligned}\quad (49.7)$$

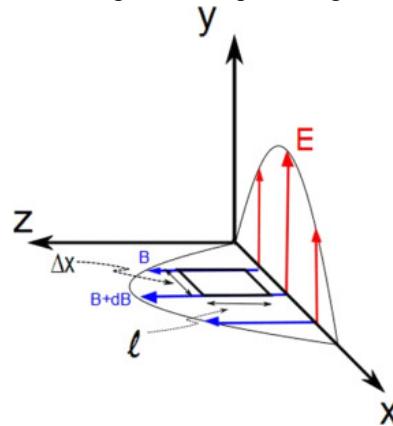
We have made some progress, we have a differential equation relating the fields, but it is a mixed equation containing both the electric and magnetic fields. We are only half way there.

Rewriting of the Maxwell-Ampere Law

We have used one field equation so far and that took us part of the way. We have the Maxwell-Ampere law as well. We can use this to modify our result from Faraday's law to find the linear wave equation that we expect. The Maxwell-Ampere law with no sources (charges or currents) states

$$\oint \vec{B} \cdot d\vec{s} = \epsilon_0 \mu_0 \frac{d\Phi_E}{dt}$$

This time we must consider the magnetic field path integral



We can do the same thing we did with Faraday's law with an area, but this time we will use the area within the magnetic field (shown in the figure above). Again, let's start with the left hand side of the equation. We see that the sides of our area that are parallel to the x -axis do not matter because $\vec{B} \cdot d\vec{s} = 0$ along these sides, but the other two are in the direction (or opposite direction) of the field. They do contribute to the line integral.

$$\begin{aligned}\oint \vec{B} \cdot d\vec{s} &= B(x, t)\ell - B(x + \Delta x, t)\ell \\ &\approx -\ell \frac{\partial B}{\partial x} \Delta x\end{aligned}\quad (49.8)$$

Now for the left hand side, we need the electric flux. For such a small area, the field is

nearly constant so

$$\begin{aligned}\Phi_E &\approx EA \cos \theta \\ &= EA \\ &= E\ell \Delta x\end{aligned}$$

so

$$\frac{\partial \Phi_E}{\partial t} = \ell \Delta x \frac{\partial E}{\partial t} \quad (49.9)$$

Combining both sides

$$\begin{aligned}\oint \vec{B} \cdot d\vec{s} &= \varepsilon_0 \mu_0 \frac{d\Phi_E}{dt} \\ -\ell \frac{\partial B}{\partial x} \Delta x &= \varepsilon_0 \mu_0 \ell \Delta x \frac{\partial E}{\partial t} \\ \frac{\partial B}{\partial x} &= -\varepsilon_0 \mu_0 \frac{\partial E}{\partial t}\end{aligned} \quad (49.10)$$

We now have a second differential equation relating B and E . But it is also a mixed differential equation.

Wave equation for plane waves

This leaves us with two equations to work with

$$\frac{\partial E}{\partial x} = -\frac{\partial B}{\partial t} \quad (49.11)$$

$$\frac{\partial B}{\partial x} = -\varepsilon_0 \mu_0 \frac{\partial E}{\partial t} \quad (49.12)$$

Remember that these are all partial derivatives. Taking the derivative of the first equation with respect to x gives

$$\begin{aligned}\frac{\partial}{\partial x} \frac{\partial E}{\partial x} &= \frac{\partial}{\partial x} \left(-\frac{\partial B}{\partial t} \right) \\ \frac{\partial^2 E}{\partial x^2} &= -\frac{\partial}{\partial x} \left(\frac{\partial}{\partial t} B \right) \\ \frac{\partial^2 E}{\partial x^2} &= -\frac{\partial}{\partial t} \left(\frac{\partial B}{\partial x} \right)\end{aligned}$$

In the last equation we swapped the order of differentiation for the right hand side. In parenthesis, we have $\partial B / \partial x$ on the right hand side. But we know what $\partial B / \partial x$ is from our second equation. We substitute from our second equation to obtain

$$\begin{aligned}\frac{\partial^2 E}{\partial x^2} &= -\frac{\partial}{\partial t} \left(-\varepsilon_0 \mu_0 \frac{\partial E}{\partial t} \right) \\ \frac{\partial^2 E}{\partial x^2} &= \varepsilon_0 \mu_0 \frac{\partial^2 E}{\partial t^2}\end{aligned} \quad (49.13)$$

We can do the same thing, but taking derivatives with respect to time to give

$$\frac{\partial^2 B}{\partial x^2} = \varepsilon_0 \mu_0 \frac{\partial^2 B}{\partial t^2} \quad (49.14)$$

You will recognize both of these last equations as being in the form of the linear wave equation.

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2}$$

This means that both the E field and the B field are governed by the linear wave equation with the speed of the waves given by

$$v = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \quad (49.15)$$

We have studied waves, so we know the solution to this equation is a sine or cosine function

$$E = E_{\max} \cos(kx - \omega t) \quad (49.16)$$

$$B = B_{\max} \cos(kx - \omega t) \quad (49.17)$$

with

$$k = \frac{2\pi}{\lambda}$$

and

$$\omega = 2\pi f$$

then

$$\frac{\omega}{k} = \frac{2\pi f}{\frac{2\pi}{\lambda}} = \lambda f$$

which is the wave speed.

We can show that the magnitude of E is related to B .

Lets take derivatives of E and B with respect to x and t .

$$\begin{aligned} \frac{\partial E}{\partial x} &= -kE_{\max} \sin(kx - \omega t) \\ \frac{\partial B}{\partial t} &= \omega B_{\max} \sin(kx - \omega t) \end{aligned}$$

then we can use one of our half-way-point equations from above

$$\frac{\partial E}{\partial x} = -\frac{\partial B}{\partial t}$$

and by substitution obtain

$$-kE_{\max} \sin(kx - \omega t) = -\omega B_{\max} \sin(kx - \omega t)$$

$$-kE_{\max} = -\omega B_{\max}$$

or

$$\frac{E_{\max}}{B_{\max}} = \frac{\omega}{k} = v$$

The speed is the speed of light, c , so

$$\frac{E_{\max}}{B_{\max}} = c \quad (49.18)$$

It is one of the odd things about the universe that speed of electromagnetic waves is a constant. It does not vary in vacuum, and the in-vacuum value, c is the maximum speed. It was a combination of Maxwell's work in predicting c and the observations confirming the predictions that launched Einstein to form the Special Theory of Relativity!

Note that the last equation shows why we often only deal with the electric field wave when we do optics. Since the magnetic field is proportional to the electric field, we can always find it from the electric field.

Properties of EM waves

Pick up here

Knowing that the electric and magnetic fields form plane waves, we can investigate these plane wave solutions to see what they imply.

Energy in an EM wave

The electromagnetic (EM) wave is a wave. Waves transfer energy. It is customary find a vector that describes the flow of energy in the electromagnetic wave. This is like the ray vectors we have been drawing for some time, but with the magnitude of the vector giving the energy flow rate.

The rate of at which energy travels with the EM wave is given the symbol \mathbf{S} and is called the Poynting vector after the person who thought of it. It is

$$\mathbf{S} = \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} \quad (49.19)$$

Let's deal with a dumb name first: The Poynting vector. It is named after a scientist with the last name Poynting. The name is really meaningless. There is nothing particularly "pointy" about this vector more than any other vector.

Instead of a formal derivation, let's just see what we get from Poynting's equation for a plane wave.

For our plane wave case, E and B are at 90° angles³⁸. so

$$S = \frac{1}{\mu_0} EB \quad (49.20)$$

³⁸ For other fields this might not be true, but it is generally true for light.

and S will be perpendicular to both. Notice from our preceding figures that this is also the direction that the wave travels! That is comforting. That should be true for a EM wave. The energy, indeed, goes the way the Poynting vector points.

Using

$$\frac{E}{B} = c$$

we can write the magnitude of the Poynting vector as

$$S = \frac{E^2}{c\mu_0} \quad (49.21)$$

We could also express this in terms of B only.

You will remember that our eyes don't track the oscillations of the electromagnetic waves. Few detectors (if any) can. For visible light, the frequency is very high. We usually see a time average. This time average of the Poynting vector is called the *intensity* of the wave

$$I = S_{ave}$$

Intensity of the waves

When we studied waves, we learned that waves have an intensity. The intensity of electromagnetic waves must relate to the strength of the fields. We can write it as

$$I = \frac{EB}{2\mu_0}$$

(can you remember where the "1/2" came from?)³⁹. Again using

$$E = cB$$

we can write the intensity as

$$I = \frac{1}{2\mu_0 c} E^2 \quad (49.22)$$

We remember that I is proportional to the square of the maximum electric field strength from our previous consideration of light intensity. But before we only said that it was proportional. Now we know the constant of proportionality. Of course we could also write the intensity as

$$I = \frac{c}{2\mu_0} B^2 \quad (49.23)$$

but this is less traditional. We have said already that the intensity, I , is the magnitude of the average Poynting vector S_{ave} .

³⁹ This is because the average value of $\sin^2(\omega t)$ over a period is given by $\frac{1}{T} \int_0^T \sin^2(\omega t) dt = \frac{\omega}{2\pi} \int_0^{2\pi/\omega} \sin^2(\omega t) dt = \frac{1}{2}$

Recall that we know the energy densities in the fields

$$\begin{aligned} u_E &= \frac{1}{2}\epsilon_0 E^2 \\ u_B &= \frac{1}{2}\frac{B^2}{\mu_0} \end{aligned}$$

again, since

$$E = cB \quad (49.24)$$

we can write

$$\begin{aligned} u_B &= \frac{1}{2}\frac{B^2}{\mu_0} \\ &= \frac{1}{2}\frac{E^2}{c^2\mu_0} \\ &= \frac{1}{2}\epsilon_0 E^2 \end{aligned} \quad (49.25)$$

so for a plane electromagnetic wave

$$u_E = u_B \quad (49.26)$$

The total energy in the field is just the sum

$$u = u_E + u_B = \epsilon_0 E^2 \quad (49.27)$$

But when we do the time average to find the intensity, we pick up a factor of a half

$$u_{ave} = \frac{1}{2}\epsilon_0 E^2 \quad (49.28)$$

Comparing this to our equation for intensity gives

$$I = \frac{1}{2\mu_0 c} E_{\max}^2 = S_{ave}$$

and then

$$\begin{aligned} S_{ave} &= \frac{1}{\epsilon_0 \mu_0 c} \frac{1}{2} \epsilon_0 E^2 \\ &= \frac{1}{\epsilon_0 \mu_0 c} u_{ave} \\ &= \frac{1}{\frac{1}{c^2} c} u_{ave} \\ &= cu_{ave} \end{aligned} \quad (49.29)$$

If you have already taken your course on thermodynamics you, learned that we could transfer energy by radiation. This is our radiation! And we see that it does indeed transfer energy. We learned about this by discussing solar heating and by talking about Army weapons that apply energy to crowds.



US Army Active Denial System (ADS).

https://jnlwp.defense.gov/Portals/50/Documents/Resources/Presentations/Joint_Integration_Program_Advanced_Planning/08-25-125259-783

but we really use this every day when we microwave something. Microwaves are electromagnetic waves!

Momentum of light

One of the strangest things is that there is also momentum in the electromagnetic waves.

If the waves are absorbed, the momentum is

$$p = \frac{U}{c} \quad (49.30)$$

or if the waves are reflected it is

$$p = \frac{2U}{c} \quad (49.31)$$

(think of balls bouncing off a wall, the change in momentum is always $2mv$ for a bounce).

We can think of the light exerting a pressure on the surface. Force is given by

$$\begin{aligned} F &= ma \\ &= m \frac{dv}{dt} \\ &= \frac{dp}{dt} \end{aligned}$$

then using this force, the pressure is

$$P = \frac{F}{A} = \frac{1}{A} \frac{dp}{dt} \quad (49.32)$$

then

$$P = \frac{F}{A} = \frac{1}{cA} \frac{dU}{dt} \quad (49.33)$$

We found $\frac{1}{A} \frac{dU}{dt}$ to be the energy rate per unit area, which is the magnitude of the Poynting vector, S . So our pressure due to light is

$$P = \frac{S}{c} \quad (49.34)$$

for perfect absorption. If there is perfect reflection

$$P = \frac{2S}{c} \quad (49.35)$$

This may seem a little strange. Water or sound waves would exert a pressure because the water or air particles can strike a surface, exerting a force. But remember the electromagnetic fields will create forces on the electrons in atoms⁴⁰, and most of the electrons are bound to the atoms in materials by the Coulomb force. So there really is a force on the material due to the electromagnetic wave. Quantum mechanics tells us about electrons being knocked out of shells into higher energy shells (absorbing photons of light) and re-emitting the light when the electrons fall back down to lower shells. This is a little like catching a frisbee, and then throwing it. Momentum is transferred both at the catch and at the release.

A cool use of this phenomena is called laser levitation

⁴⁰ Protons too, but the protons are more tightly bound due to the nuclear strong force and the nuclei are bound in the material. Their resonant frequencies are usually not assessable to visible light, so I will ignore their effect in our treatment. But if you consider x-rays or gamma rays, they would be important.



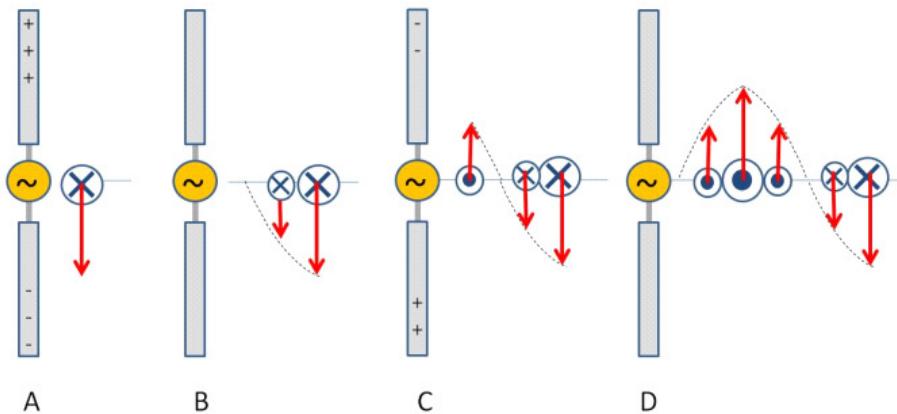
Laser Levitation (Skigh Lewis, Larry Baxter, Justin Peatross (BYU), Laser Levitation: Determination of Particle Reactivity, ACERC Conference Presentation, February 17, 2005)

In the picture you are seeing a single small particle that is floating on a laser beam. the laser beam is directed upward. The force due to gravity would make the particle fall, but the laser light keeps it up!

Antennas Revisited

We talked about antennas before. Let's try to put all we have done together to make a radio wave. First, we know from our analysis that we need changing fields. Neither static charges, nor constant currents will do. If we think about this for a minute, we will realize that the charges will *accelerate*. Fundamentally, this is the mechanism for making EM waves.

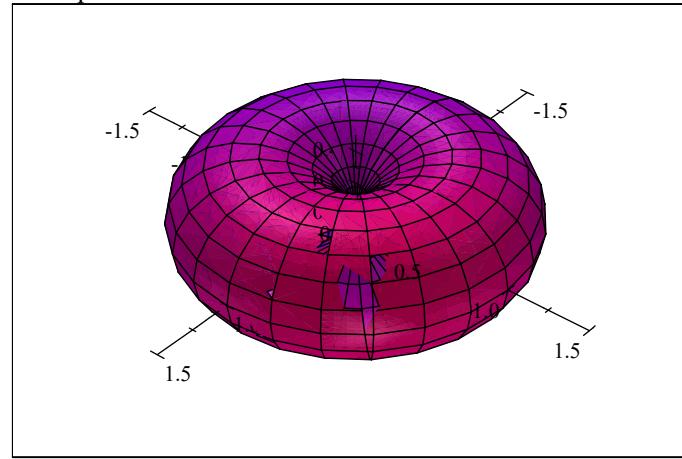
The half wave antenna is simple to understand, so let's take it as our example.



It is made from two long wires connected to an alternating current source (the radio transmitter). The charges are separated in the antenna as shown. But the separation switches as the alternating current changes direction. The charges accelerate back and forth, like a dipole switching direction. Radio people call this antenna a *simple dipole*.

Note the direction of the E and B fields. The Poynting vector is to the right. The antenna field sets up a situation far from the antenna, itself, where the changing electric field continually induces a magnetic field and the changing magnetic field continually induces a changing electric field. The wave becomes self sustaining! And the energy it carries travels outward.

Below you can see a graph of the sort of toroidal angular dependence of the dipole antenna emission pattern.



Angular dependence of S for a dipole scatterer.

From this you can see why we usually stand antennas straight up and down. Then the

transmission travels parallel to the Earth's surface, where receivers are more likely to be.

Speaking of receivers, of course the receiver works like a transmitter, only backwards. The EM waves that hit the receiving antenna accelerate the electrons in the wire of the antenna. The induced current passed through an LRC circuit who's resonance frequency allows amplification of just one small band of frequencies (the one your favorite radio station is using) and then the amplified signal is sent to a speaker.

The Electromagnetic Spectrum

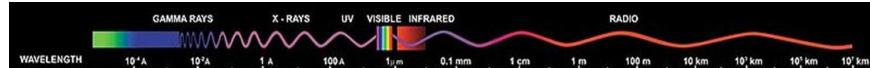
Maxwell predicted how fast his field waves would travel by finding the linear wave equation from the fields and noticing the speed indicated by the result. We have seen how he did this. The answer is

$$v = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \quad (49.36)$$

this speed is so special in physics that it get's its own letter

$$c = 2.99792 \times 10^8 \frac{\text{m}}{\text{s}} \quad (49.37)$$

which is of course the speed of light. In fact, that this was the measured speed of light was strong evidence leading us to conclude that light was really a type of these waves. There are a few more types of electromagnetic waves. In the following chart you can see that visible light is just a small part of what we call the *electromagnetic spectrum*.



Electromagnetic Spectrum (Public Domain image courtesy NASA)

The speed of light is always a constant in vacuum. This is strange. It caused a lot of problems when it was discovered.

$$v = f\lambda \quad (49.38)$$

or

$$c = f\lambda \quad (49.39)$$

where we can see that for light and electromagnetic waves, knowing the wavelength is always enough to know the frequency as well (in a vacuum).

As an example of what problems can come, let's consider a Doppler effect for light. Remember for sound waves, we had a Doppler effect. We will have a Doppler effect for electromagnetic waves too. But light does not change it's speed relative to a reference

frame. This is *really weird*. The speed of light in a vacuum is *always c—no matter what frame we measure it in*.

Einstein's theory of Special Relativity is required to deal with this constant speed of light in every reference frame. From Relativity, the Doppler equation is

$$f' = f \frac{\sqrt{1 + \frac{v}{c}}}{\sqrt{1 - \frac{v}{c}}} \quad (49.40)$$

or, if we let u be the relative velocity between the source and the detector, and insist that $u \ll c$

$$f' = f \left(\frac{c + u}{c} \right) \quad (49.41)$$

Where of course f' is the observed frequency and f is the frequency emitted by the source. This is usually written as

$$f' = f \left(1 \pm \frac{u}{c} \right) \quad (49.42)$$

but it is really the same equation⁴¹. Just like with sound, we use the positive sign when the source and observer are approaching each other.

This means that if things are moving closer to each other the frequency increases. Think of

$$\lambda = \frac{c}{f} \quad (49.43)$$

this means that as a source and emitter approach each other, then the light will have a shorter wavelength. Think of our chart on the electromagnetic spectrum. This means the light will get bluer. If they move farther apart, the light will get redder.

This is what gave us the hint that has lead to our cosmological theories like the big bang. Although this theory is now much more complicated, the facts are that as we look at far away objects, we see they are all *red shifted*. That is, they all show absorption spectra for known elements, but at longer wavelengths than we expect from laboratory experiments. We interpret this as meaning they are all going away from us!

Summary

Here is what we have learned so far about the properties of light

1. Electromagnetic waves travel at the speed of light
2. Electromagnetic waves are transverse electric and magnetic waves that are oriented perpendicular to each other.

⁴¹ This equation is only really true for relative speeds u that are much less than the speed of light. Since it is very hard to make something travel even close to the speed of light, we will find it is nearly always true.

3. $E = cB$
4. Electromagnetic waves carry energy *and momentum*

Photons

Our understanding of light is not complete yet. If you went on to take PH279 you would find that light still operates much like a particle at times. This should not be a surprise, since Newton and others explained much of optics (the study of light) assuming light was a particle.

Einstein and others noticed that for some metals, light would strike the surface and electrons would leave the surface. The energy of a wave is proportional to the amplitude of the wave. It was expected that if the amplitude of the electromagnetic wave was increased, the number of electrons leaving the surface would increase. This proved to be true most of the time. But Hertz and others decided to try different frequencies of light. It turns out that as you lower the frequency, all of a sudden no electrons leave no matter how big the amplitude of the wave. Something was wrong with our wave theory of light. The answer came from Einstein who used the idea of a “packet” of light to explain this *photoelectric effect*. For now, we should know just that the waves of light exist in *quantized* packets called *photons*. The energy of a photon is

$$E = hf \quad (49.44)$$

where E is the energy, f is the frequency of the light wave, and h is a constant

$$h = 6.63 \times 10^{-34} \text{ J s} \quad (49.45)$$

A beam of light is many, many photons all superimposing. We know how waves combine using superposition, so it is easy to see that we can get a big wave from many little waves.

Knowing that light is made from electric and magnetic fields, and that these fields are vector fields, we should expect some directional quality in light. And there is such a directional quality that we will study next lecture.

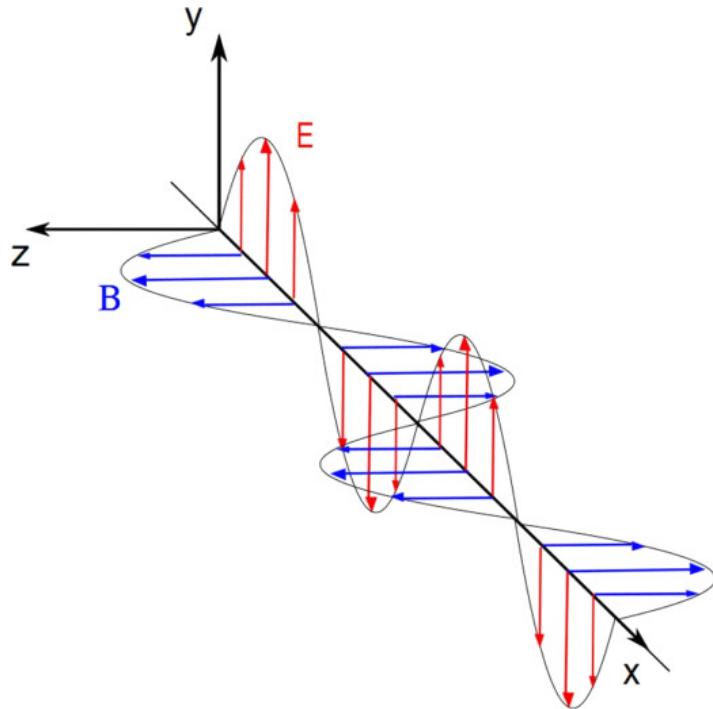
50 Polarization

Fundamental Concepts

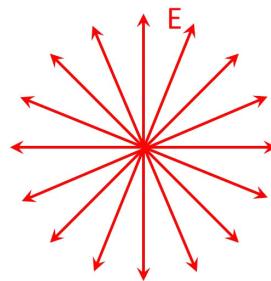
- The direction of the electric field in a plane wave is called the polarization direction.
- Natural light is usually a superposition of many waves with random polarization directions. This light is called unpolarized light.
- Some materials allow light with one polarization to pass through, while stopping other polarizations. The polaroid is one such material polaroids. will have a final intensity that follows the relationship $I = I_{\max} \cos^2(\theta)$
- Light reflecting off a surface may be polarized because of the absorption and re-emission pattern of light interacting with the material atoms.
- Scattered light may be polarized because of anisotropies in the scatterers.
- Birefringent materials have different wave speeds in different directions. This affects the polarization of light entering these materials.

Polarization of Light Waves

We said much earlier in our study of light that it was a transverse wave. Last lecture we saw that we have an electric and magnetic field direction, and that these directions are perpendicular to each other and the direction of energy flow. We will now show some implications of this fact. In a course in electromagnetic theory, we often draw light as in the figure below.



We will continue to ignore the magnetic field (marked in the figure as B). We will look at the E field and notice that it goes up and down in the figure. But we could have light in any orientation. If we look directly at an approaching beam of light we would “see” many different orientations as shown in the next figure.



When light beams have waves with many orientations, we say they are *unpolarized*. But suppose we were able to align all the light so that all the waves in the beam were transverse waves in the same orientation. Say, the one in the next figure.

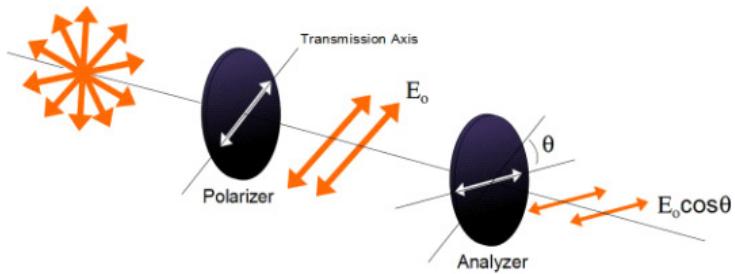


Then we would describe the light as *linearly polarized*. The plane that contains the E -field is known as the *polarization plane*.

Polarization by removing all but one wave orientation

One way to make polarized light is to remove all but one orientation of an unpolarized beam. A material that does this at visible wavelengths is called a *polaroid*. It is made of long-chain hydrocarbons that have been treated with iodine to make them conductive. The molecules are all oriented in one direction by stretching during the manufacturing process. The molecules have electrons that can move when light hits them. They can move farther in the long direction of the molecule, so in this direction the molecules act like little antennas. The molecules' electrons are driven into harmonic motion along the length of the molecule. This takes energy (and therefore, light) out of the beam. Little electron motion is possible in the short direction of the molecule, so light is given a preferential orientation. The light is passed if it is perpendicular to the long direction of the molecules. This direction is called the *transmission axis*.

We can take two pieces of polaroid material to study polarization.



Unpolarized light is initially polarized by the first piece of polaroid called the *polarizer*. The second piece of polaroid then receives the light. This piece is called the *analyzer*. If

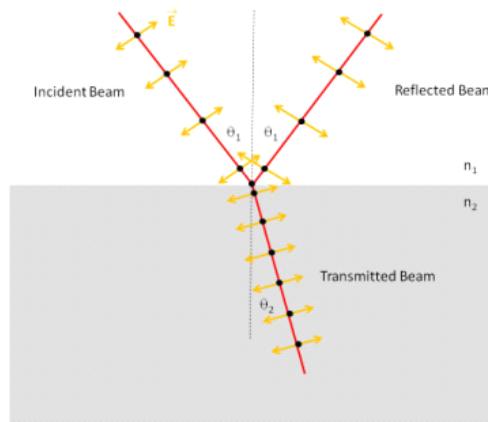
there is an angular difference in the orientation of the transmission axes of the polarizer and analyzer, there will be a reduction of light through the system. We expect that if the transmission axes are separated by 90° no light will be seen. If they are separated by 0° , then there will be a maximum. It is not hard to believe that the intensity will be given by

$$I = I_{\max} \cos^2(\theta) \quad (50.1)$$

remembering that we must have a squared term because $I \propto E^2$.

Polarization by reflection

If we look at light reflected off of a desk or table through a piece of polaroid, we can see that at some angles of orientation, the reflection diminishes or even disappears! Light is often polarized on reflection. Let's consider a beam of light made of just two polarizations. We will define a plane of incidence. This plane is the plane of the paper or computer screen. This plane is perpendicular to the reflective or refractive surface in the figure below.

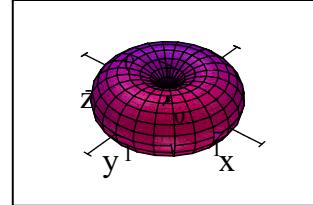


One of our polarizations is defined as parallel to this plane. This direction is represented by orange (lighter grey in black and white) arrows in the figure. The other polarization is perpendicular to the plane of incidence (the plane of the paper). This is represented by the black dots in the figure. These dots are supposed to look like arrows coming out of the paper.

When the light reaches the interface between n_1 and n_2 it drives the electrons in the medium into SHM. The perpendicular polarization finds electrons that are free to move in the perpendicular direction and re-radiate in that direction. Even for a dielectric, the

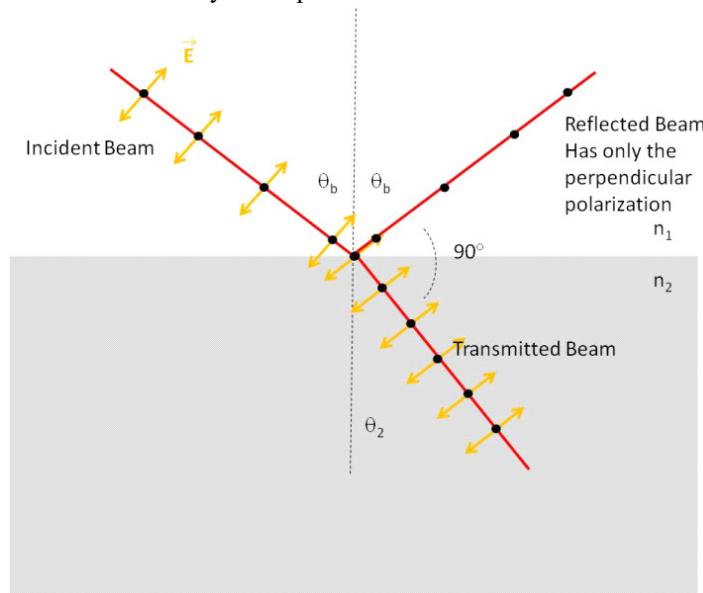
electron orbitals change shape and oscillate with the incoming electromagnetic wave.

The parallel ray is also able to excite SHM, but a electromagnetic analysis tells us that these little “antennas” will not radiate at an angle 90° from their excitation direction. Think of little dipole radiators. We can plot the amplitude of the electric field as a function of direction around the antenna.



Angular dependence of S for a dipole scatterer.

We see that along the antenna axis, the field amplitude is zero. This means that the wave really does not go that direction. So in our case, the amount of polarization in the parallel direction decreases with the angle between the reflected and refracted rays until at 90° there is no reflected ray in the parallel direction.



The incidence angle that creates an angular difference between the refracted and reflected rays of 90° is called the Brewster's angle after its discoverer. At this angle the reflected beam will be completely linearly polarized.

We can predict this angle. Remember Snell's law.

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

Let's re-lable the incidence angle $\theta_1 = \theta_b$. We take $n_1 = 1$ and $n_2 = n$ so

$$n = \frac{\sin \theta_b}{\sin \theta_2}$$

Now notice that for Brewster's angle, we have

$$\theta_b + 90^\circ + \theta_2 = 180^\circ$$

so

$$\theta_2 = 90^\circ - \theta_b$$

so we have

$$n = \frac{\sin \theta_b}{\sin (90^\circ - \theta_b)}$$

ah, but we remember that $\sin (90^\circ - \theta) = \cos (\theta)$ so

$$n = \frac{\sin \theta_b}{\cos \theta_b}$$

but again we remember that

$$\tan \theta = \frac{\sin \theta}{\cos \theta}$$

so

$$n = \tan \theta_b \quad (50.2)$$

which we can solve for θ_b .

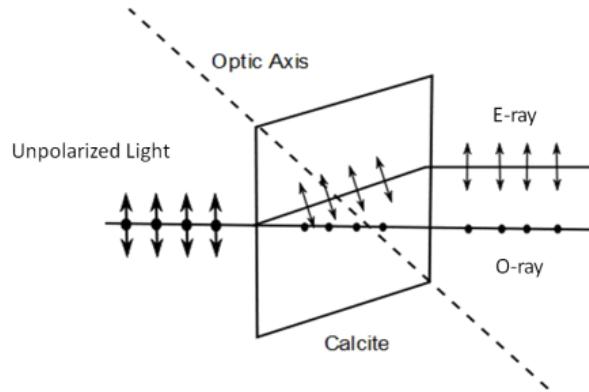
$$\theta_b = \tan^{-1} (n)$$

This phenomena is why we wear polarizing sunglasses to reduce glare.

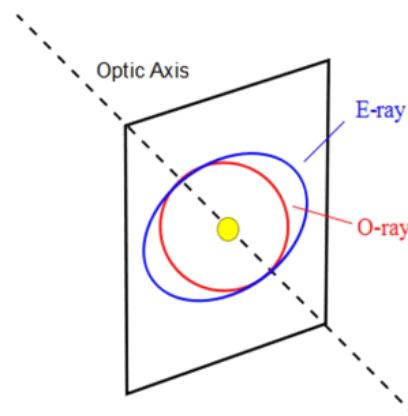
Birefringence

Glass is an amorphic solid—that is—it has no crystal structure to speak of. But some minerals do have definite order. Sometimes the difference in the crystal structure creates a difference in the speed of propagation of light in the crystal. This is not to hard to believe. We said before that the reason light slows down in a substance is because it encounters atoms which absorb and re-emit the light. If there are more atoms in one direction than another in a crystal, it makes sense that there could be a different speed in each direction.

Calcite crystals exhibit this phenomena. We can describe what happens by defining two polarizations. One parallel to the plane of the figure below, and one perpendicular.



With a careful setup, we can arrange things so the perpendicular ray is propagated just as we would expect for glass. We call this the *O*-ray (for *ordinary*). The second ray is polarized parallel to the incidence plane. It will have a different speed, and therefore a different index of refraction. We call it the *Extraordinary ray* or *E*-ray.



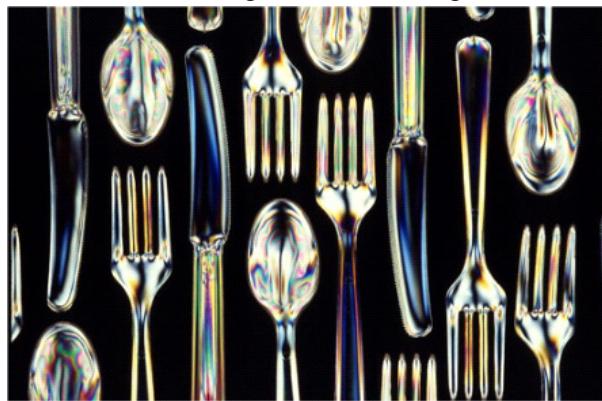
If we were to put a light source in a calcite crystal, we would see the *O*-ray send out a sphere of light as shown in the figure above. But the *E*-ray would send out an ellipse. The speed for the *E*-ray depends on orientation. There is one direction where the speeds are equal. This direction is called the *optic axis* of the crystal.



If our light entering our calcite crystal is unpolarized, then we will have two images leaving the other side that are slightly offset because the *O*-rays and *E*-rays both form images.

Optical Stress Analysis

Some materials (notably plastics) become birefringent under stress. A plastic or other stress birefringent material is molded in the form planned for a building or other object (usually made to scale). The model is placed under a stress, and the system is placed between two polaroids. When unstressed, no light is seen, but under stress, the model changes the polarization state of the light, and bands of light are seen.



Polarization due to scattering

It is important to understand that light is also polarized by scattering. It really takes a bit of electromagnetic theory to describe this. So for a moment, let's just comment that blue light is scattered more than red light. In fact, the relative intensity of scattered light goes like $1/\lambda^4$. This has nothing to do with polarization, but it is nice to know.

Now suppose we have long pieces of wire in the air, say, a few microns long. The pieces of wire would have electrons that could be driven into SHM when light hits them. If the wires were all oriented in a common direction, we would expect light to be absorbed if it was polarized in the long direction of the particles and not absorbed in a direction perpendicular to the orientation of the particles. This is exactly what happens when long ice particles in the atmosphere orient in the wind (think of the moment of inertia).

We often get impressive halo's around the sun due to scattering from ice particles.

Rain drops also have a preferential scattering direction because they are shaped like oblate spheroids (not "rain drop shape" like we were told in grade school).

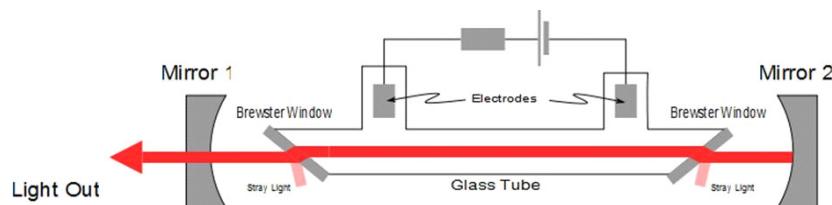
It is also true that small molecules will act like tiny antennas and will scatter light preferentially in some directions and not in others. This is called *Rayleigh scattering* and is very like small dipole antennas.

Optical Activity

Some substances will rotate the polarization of a beam of light. This is called being *optically active*. The polarization state of the light exiting the material depends on the length of the path through the material. Your calculator display works this way. An electric field changes the optical activity of the liquid crystal. There are polarizers over the liquid crystal, so sometimes light passes through the display and sometimes it is black.

Laser polarization

One last comment. Lasers are usually polarized. This is because the laser light is generated in a *cavity* created by two mirrors. The mirror is tipped so light approaches it at the Brewster angle. Light with the right polarization (parallel to the plane of the drawing) is reflected back nearly completely, but light with the opposite polarization is not reflected at all. This reduces the usual loss in reflection from a mirror, because in one polarization the light must be reflected completely.



Retrospective

We have thought about many things in this class. It has been a class *about* science. It has not been a class where we have tried to discover new science, or practiced the scientific method. This is on purpose, this being an engineering class designed to teach the principles of physics for use in designing machines.

But we should pause to think, just for a moment, about the philosophy of science. Is everything in these lectures true? We did not perform experiments to show every principle we learned. So does it all work?

The answer is—maybe. Experiments have been done to show that the equations we have learned work at least sometimes. But science is an inductive process. We can't prove anything true with science. We can only prove things false. So what we have studied is what has not been proven false, yet. Of course, even then, we have taken approximations from time to time, but we pointed these out along the way. You will know when the approximations will fail, because we talked about their valid ranges.

It is important to remember that we are not done discovering new things, and proving old things false. The laws of Newton are approximations that work at low speeds. Relativity provides mechanical equations for very high speeds (e.g. the satellite motion involved in the GPS system). But is Relativity correct? We think it works pretty well, but really we don't know. We may never know for sure. But we know it works within the range of things we have tried.

There are physicists today that are working on a fundamentally new model of the universe. It is called “String Theory” and it would replace most of our thoughts about how matter is made and how it interacts. The equations would reduce to the ones we used in class for the conditions we considered. That is because the new equations have to match the results of the experiments that we have already done or they can't be correct. But the explanations might be very different.

Often, it is in using physics to build something that we learn about the limitations of physical theory. You may be part of that process. It is a happy process because extending our understanding allows us to build new things. But don't be surprised if some of the things we learned in this class are different by the time your children take their engineering physics course. That is what we should expect of an inductive process.

It is also important to note that revealed truth is not an inductive process. It is still not

static (see article of faith 9), but it *can* prove something true as well as prove things false. I hope your FDSCI 101 experience gave you some insight into doing science as well as learning about science.

Some members view science and revelation as in opposition. But I think they are complementary. The scientific process allows us to eliminate things that are not true, allowing us to follow D&C 9:8 in preparation for seeking revelation. During a recent convocation speech, Elder Scott described using this process as a nuclear engineer during his engineering career . We can use this combination in our personal lives as well. I hope you will consider this in your careers and lives.

I have tried to give at least equal time to conceptual understanding and mathematical solving. I hope you review and refresh the conceptual understanding of the physics of what you build. Most of my industrial career, we built what we designed very well. We always did our calculations well. But we did, at times, build the wrong thing because the conceptual basis of the design was wrong. Such mistakes are difficult to fix. Conceptual understanding is a guiding principle for a successful design career. I hope this class has contributed to that conceptual understanding.

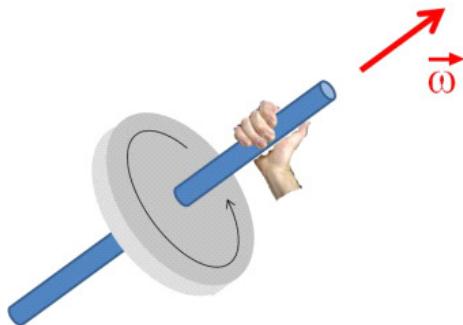
51 Summary of Right Hand Rules

PH121 or Dynamics Right Hand Rules

We had two right hand rules on PH121. We didn't give them numbers back then, so we will do that now.

Right hand rule #0:

We found that angular velocity had a direction that was given by imagining you grab the axis of rotation with your right hand so that your fingers seem to curl the same way the object is rotating. Then your thumb gives the direction of $\vec{\omega}$

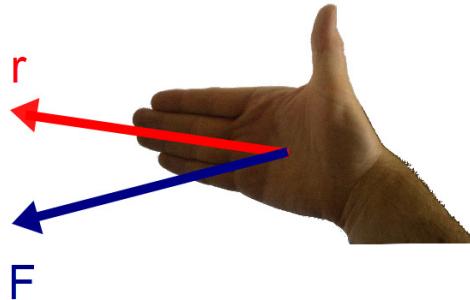


You curl the fingers of your right hand (sorry left handed people, you have to use your right hand for this) in the direction of rotation. Then your thumb points in the direction of the vector.

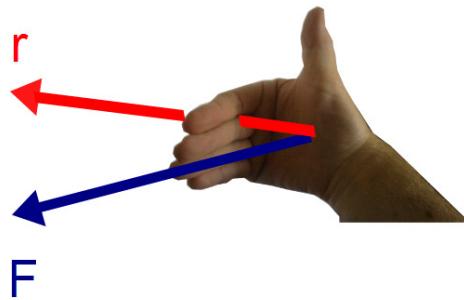
Right hand rule #0.5:

To find the direction of torque, we used the following procedure

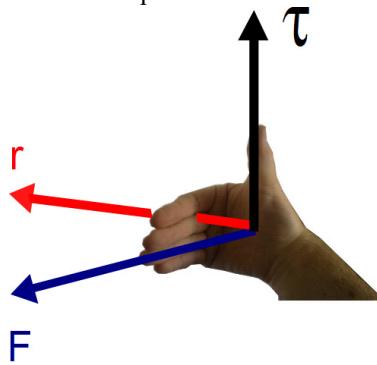
1. Put your fingers of your right hand in the direction of \tilde{r}



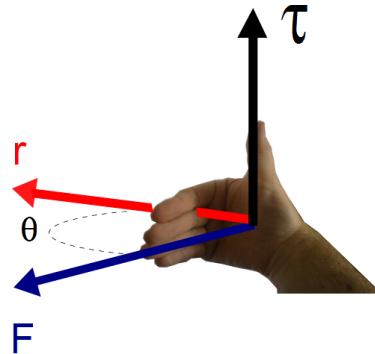
2. Curl them toward \tilde{F}



3. The direction of your thumb is the torque direction



4. The angle θ is the angle between \tilde{r} and \tilde{F}



The magnitude of the torque is

$$\tau = rF \sin \theta$$

PH223 Right Hand Rules

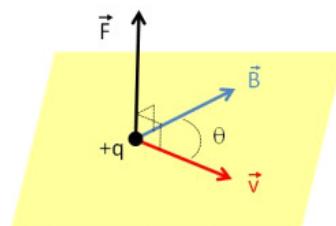
We have four more right hand rules this semester having to do with charges and fields.

Right hand rule #1:

From this rule we get the **direction of the force on a moving charged particle** as it travels thorough a **magnetic field**.

This rule is very like torque. We start with our hand pointing in the direction of \vec{v} . Curl your fingers in the direction of \vec{B} . And your thumb will point in the direction of the force. The magnitude of the force is given by

$$F = qvB \sin \theta \quad (51.1)$$



Right hand rule #2:

From this rule we get the direction of the force on current carrying wire that is in a magnetic field.

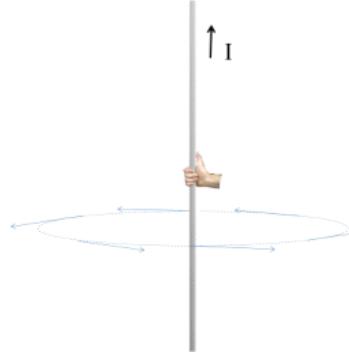
This rule is very like right hand rule #1 above. We start with our hand pointing in the direction of \mathbf{I} . Curl your fingers in the direction of $\tilde{\mathbf{B}}$. And your thumb will point in the direction of the force. The magnitude of the force is given by

$$F = ILB \sin \theta \quad (51.2)$$

Right hand rule #3:

From this rule we get the **direction of the magnetic field that surrounds a long current carrying wire**.

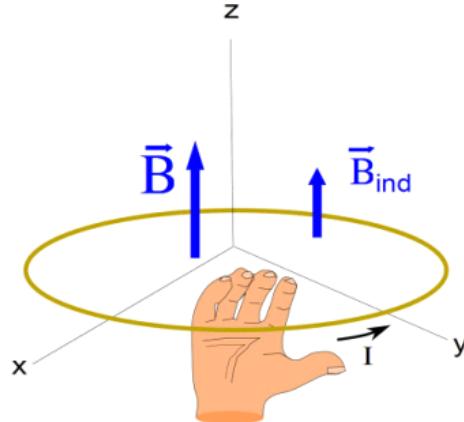
This rule is quite different. It is reminiscent of the rule for angular velocity, but there are some major differences as well. The field is a magnitude and a direction at every point in space. We can envision drawing surfaces of constant field strength. They will form concentric circles (really cylinders) centered on the wire. At any one point on the circle the field direction will be along a tangent to the circle. The direction of the vector is given by imaging you grab the wire with your right hand (don't really do it). Grab such that your right thumb is in the direction of the current. Your fingers will naturally curl in the direction of the field.

**Right Hand Rule #4:**

From this rule we get the **direction of the induced current when a loop is in a**

changing magnetic field.

This rule is only used when we have a loop with a changing external magnetic field. The rule gives the direction of the induced current. The induced magnetic field will oppose the change in the external field, trying to prevent a change in the flux. The current direction is found by imagining we stick our right hand into the loop in the direction of the induced field. Keeping our hand inside the loop we grab a side of the loop. The current goes in the direction indicated by our thumb.



In the figure above, the external field is upward but decreasing. So the induced field is upward. The current flows because there is an induced *emf* given by

$$\begin{aligned}\mathcal{E} &= -N \frac{\Delta\Phi}{\Delta t} \\ &= -N \frac{(B_2 A_2 \cos \theta_2 - B_1 A_1 \cos \theta_1)}{\Delta t}\end{aligned}$$

52 Some Helpful Integrals

$$\int \frac{r dr}{\sqrt{r^2 + x^2}} = \sqrt{r^2 + x^2}$$

$$\begin{aligned}\int \frac{dx}{(x^2 \pm a^2)^{\frac{3}{2}}} &= \frac{\pm x}{a^2 \sqrt{x^2 \pm a^2}} \\ \int \frac{x dx}{(x^2 \pm a^2)^{\frac{3}{2}}} &= \frac{-1}{\sqrt{x^2 \pm a^2}}\end{aligned}$$

$$\int \frac{dx}{x} = \ln x$$

$$\begin{aligned}\int \frac{dx}{x^2} &= -\frac{1}{x} \\ \int_0^{2\pi} \int_0^\pi \int_0^R \sin \theta d\theta d\phi &= 4\pi \\ \int_0^{2\pi} \int_0^\pi \int_0^R r^2 dr \sin \theta d\theta d\phi &= \frac{4}{3}\pi R^3 \\ \int_0^{2\pi} \int_0^R r dr d\phi &= \pi R^2\end{aligned}$$

53 Table of Physical Constants

Charge and mass of elementary particles

Proton Mass	$m_p = 1.6726231 \times 10^{-27} \text{ kg}$
Neutron Mass	$m_n = 1.6749286 \times 10^{-27} \text{ kg}$
Electron Mass	$m_e = 9.1093897 \times 10^{-31} \text{ kg}$
Electron Charge	$q_e = -1.60217733 \times 10^{-19} \text{ C}$
Proton Charge	$q_p = 1.60217733 \times 10^{-19} \text{ C}$

α -particle mass ⁴²	$m_\alpha = 6.64465675(29) \times 10^{-27} \text{ kg}$
α -particle charge	$q_\alpha = 2q_e$

Fundamental constants

Permittivity of free space	$\epsilon_0 = 8.854187817 \times 10^{-12} \frac{\text{C}^2}{\text{N m}^2}$
Permeability of free space	$\mu_0 = 4\pi \times 10^{-7} \frac{\text{T m}}{\text{A}}$
Coulomb Constant	$K = \frac{1}{4\pi\epsilon_0} = 8.98755 \times 10^9 \text{ N m}^2 \text{ C}^{-2}$
Gravitational Constant	$G = 6.67259 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$
Speed of light	$c = 2.99792458 \times 10^8 \text{ m s}^{-1}$
Avogadro's Number	$6.0221367 \times 10^{23} \text{ mol}^{-1}$
Fundamental unit of charge	$q_f = 1.60217733 \times 10^{-19} \text{ C}$

Astronomical numbers

Mass of the Earth ⁴³	$5.9726 \times 10^{24} \text{ kg}$
Mass of the Moon ⁴⁴	$0.07342 \times 10^{24} \text{ kg}$
Earth-Moon distance (mean) ⁴⁵	384400 km
Mass of the Sun ⁴⁶	$1,988,500 \times 10^{24} \text{ kg}$
Earth-Sun distance ⁴⁷	$149.6 \times 10^6 \text{ kg}$

Conductivity and resistivity of various metals

⁴² <http://physics.nist.gov/cgi-bin/cuu/Value?mal>

⁴³ <http://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html>

⁴⁴ <http://nssdc.gsfc.nasa.gov/planetary/factsheet/moonfact.html>

⁴⁵ <http://solarsystem.nasa.gov/planets/profile.cfm?Display=Facts&Object=Moon>

⁴⁶ <http://nssdc.gsfc.nasa.gov/planetary/factsheet/sunfact.html>

⁴⁷ <http://nssdc.gsfc.nasa.gov/planetary/factsheet/index.html>

Material	Conductivity ($\Omega^{-1} \text{ m}^{-1}$)	Resistivity ($\Omega \text{ m}$)	Temp. Coeff. (K^{-1})
Aluminum	3.5×10^7	2.8×10^{-8}	3.9×10^{-3}
Copper	6.0×10^7	1.7×10^{-8}	3.9×10^{-3}
Gold	4.1×10^7	2.4×10^{-8}	3.4×10^{-3}
Iron	1.0×10^7	9.7×10^{-8}	5.0×10^{-3}
Silver	6.2×10^7	1.6×10^{-8}	3.8×10^{-3}
Tungsten	1.8×10^7	5.6×10^{-8}	4.5×10^{-3}
Nichrome	6.7×10^5	1.5×10^{-6}	0.4×10^{-3}
Carbon	2.9×10^4	3.5×10^{-5}	-0.5×10^{-3}