Physics 7C

Various authors
If you find errors please e-mail: djmartin@ucdavis.edu

December 31, 2007
Version 1.1.2

Contents

List of examples

Preface

How to use these notes

The way to learn physics is by doing it and thinking about it yourself, rather than just reading about it. The goal of this text is not to teach you physics, but to tell you what the underlying principles are. To become a successful physicist this means that you need to practise taking these principles and using them to *explain* or *predict* what will happen in specific situations. You will get some practise applying these rules in the discussion/lab (DL) activities, the assigned homework and extra problems in the text. This text is also designed to reinforce the principles we expect you to get out of the labs.

These notes are *not* a complete enumeration of all problems given in the course. They are supposed to be a complete enumeration of the principles we wish you to learn, and give some guidance on how to approach specific problems to illustrate those principles. The skill we want you to learn is to be able to take a description of a physical phenomenon, figure out which principles of physics are relevant for answering the question, and then construct a logical argument that starts with those principles and gets to an answer about the phenomenon.

> Do **not** spend your time in DL looking for the answers to a particular activity.

These notes do not contain the solutions to DL activities in them, as the skill you are practising in DL is working these things out. If you spend your time reading through the notes this will mean a lot of wasted time in DL. The DLs are supposed to be fairly self-contained in introducing you to an idea, and this should not be necessary.

Have a study plan

If you want to get the most out of the DLs, one way of doing it would be to read the notes beforehand to get an idea of what is being described and the models being used. Use the DL activities to "catch" where your misconceptions are. Then go back and re-read the notes and see if your question is answered there. Finally, work through the homework to test if you understand the concept. An example "study plan" is shown below:

1. Read the section of the notes
   Get basic concepts

2. DL
   Practise applying concepts
   Identify misconceptions

3. Read the section of the notes
   Clear up misconceptions

4. Solo FNTs
   Test your ability to apply principles

5. Group FNTs (outside of class)
   Test your ability to explain physics
   Ask others about things you didn't understand
   Hint: If you cannot explain why your answer is right, or why the other person's answer is right you still don't understand the underlying concept.

6. Go to office hours
   Resolve misconceptions that you couldn't resolve as a group
   As you have tried to work them out yourself the answers will stick with you longer!

The group that you use for meeting and discussing your FNTs does not need to be your group from DL. Having a different group may be beneficial as it will give you experience explaining things to different groups of people.

How this text is organised

In writing any text, the issue of what to put in and what to leave out always arises. In a textbook the author may write additional topics that never get covered in a particular course, but because this text was written specifically

for physics 7C students may think that everything in this text is testable. However, physics 7C does change slightly from quarter to quarter. Your instructor and the DL notes should tell you which parts of the notes are relevant to various parts of the course.

The sections in the notes are organised in the following way (symbols show how these sections are marked throughout the text):

- Chapters and sections
  These are the bare bones of the course notes, and provide the basic principles we are trying to teach. Some examples will be given. The goal here is to *memorize* the principles and to *learn* how to apply those principles.

- Advanced sections*
  These go into greater depth, or clear up points of possible confusion. We do not require you to fully grasp what is in these sections. These sections are there primarily for when we have over-simplified a topic and a student thinking about things carefully may be misled, and for the student that is enthusiastic about physics.

- Applications sections†
  These sections are examples of the physics principles that you have learnt already, but applied to the real world. When applied to the real world, physics is much more interesting. They differ from the main notes in that the main notes explain principles that you have to remember, while the application sections should not be memorized.

- Appendices
  With the exception of the first appendix (a collection of physics constants), the appendices are brief outlines of subjects we expect you to know from other classes, and to give you an idea of roughly the level we expect. The appendices will not be covered in class at all, so if they contain material you are unsure of seek help in office hours or find a good textbook on the subject.

What physics 7C covers

Waves, fields and a semi-classical introduction to quantum mechanics.

Acknowledgements

The majority of this text was written over six months by Damien Martin and Emily Ashbaugh, who were at the time graduate students in physics at UC Davis. The chapter on magnetic fields (9-3) was written by physics professor Manuel Calderon de la Barca Sanchez and graduate student Daniel Phillips.

A special thanks goes to Daniel York, an undergraduate taking 7C who volunteered to edit these notes the first quarter they were used. Many spelling errors and points of confusion have been cleared up due to his diligent editing efforts.

Unit 8

Waves

Unit 8:

8-1: An introduction to waves

We begin our study of waves in this first unit of Physics 7C with an introduction to waves and then a thorough development of a model – the harmonic plane wave model – which we will then use extensively to make sense of a wide variety of wave phenomena.

In this section we will familiarise ourselves with waves by concentrating on *material waves*. These are the disturbances or vibrations of atoms or molecules of a particular substance. Any sort of ripple that you have seen, as well as sound is an example of a material wave. There are other types of waves such as light or matter waves which are not as easy to visualise, and we shall postpone detailed discussion until §9-4 and §10-1 respectively.

An important message of this entire course is that one of the distinguishing features of physics is the continual striving for general principles and simple models that can be applied to large classes of phenomena. In our study of wave phenomena we very consciously take this approach; the focus is on the model and its representation, and not on one or another of an almost unlimited number of individual phenomena associated with sound, light, TV and radio waves, microwaves, etc. Our goal is to enable you to develop a useful understanding of wave behavior that you can then apply to any phenomenon that can be modeled as a wave, whether on a quiz or the course final, or more importantly, throughout your everyday life and in your professional or scientific career.

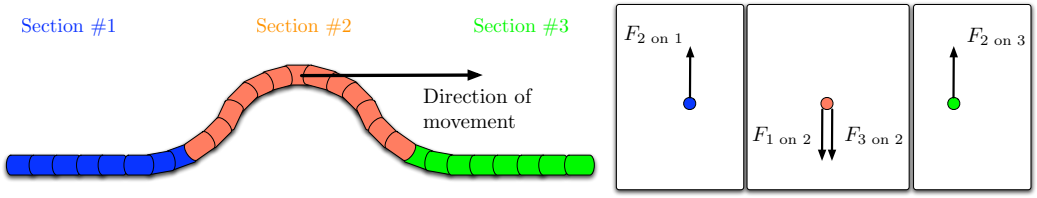
8-1-1 What is a wave?

Some primitive wave concepts

There are two important goals associated with the first part of this unit. To become familiar with wave phenomena and how we analyze them and

secondly, to understand the mathematical representation of one-dimensional harmonic waves sufficiently, so that we can use it as a tool throughout the rest of the course to help us understand the physics of sound and light (electromagnetic) waves.

A material wave is a particular, although very common, type of *internal motion* of a medium (material substance). In order for material waves to exist there must be forces between neighbouring particles in the medium. We will examine how a disturbance travels by colouring a medium three different pieces and labelling them as sections #1, #2 and #3.



The "displacement" of the medium in section #2 is pulled down by sections #1 and #3. Thus section #2 will be accelerated downward, back toward equilibrium. By Newton's third law section #2 must exert an *equal and opposite force* $\boldsymbol{F}_{2 \text{ on } 3}$ on section #3. This will cause section section #3 to accelerate upward, so a little time later section #3 will be raised. The "disturbance" has travelled from section #2 to section #3 **without** the individual pieces of medium travelling along with it! You may wonder why the sections exert a force on one another at all – the origin of this force can be traced back to the fact that the individual atoms have a preferred separation r_0 and that by stretching or squashing the medium the atoms push on their neighbours. This is the Lennard-Jones interaction that we learned about in Physics 7A! We have simply clumped atoms together into three sections for convenience – you could have the same discussion with individual atoms! These restoring forces are typically greatest in the solid phase and least in the gaseous phase.

There is one small detail that you may be wondering about – by Newton's third law there must be a upward force $\boldsymbol{F}_{2 \text{ on } 1}$. Why does the pulse not travel in *both* directions? The answer is that the pulse was previously in section #1, and to get back to the level in the picture section #1 must have travelled *down*. From Newton's first law we know that if there was no net force acting on section #1 it would keep travelling downward at a constant speed. It comes to rest precisely because there is an upward force $\boldsymbol{F}_{2 \text{ on } 1}$!

The point being made here is that the concepts we are introducing when discussing material waves are no different from the concepts that we have already introduced when discussing forces, motion and atoms. The reason we place such an emphasis on waves is that dealing with the form of a wave is often easier than trying to visualise force diagrams for many different parts of a medium. Eventually we will encounter non-material waves (such as matter waves or light) and our previous exposure to disturbances in material waves will be an invaluable guide.

The disturbance (or, more technically, an oscillation) may be spread throughout the medium and recur continuously at each point (repetitive wave) or the oscillations may exist for only a limited time at each point (pulse-type wave). The material wave that we have divided into sections is an example of a pulse-type wave.

Waves in a medium are started by outside forces that act on some of the particles in the medium to start them oscillating. External forces are not required to keep a wave going once it has been started. In the example of a surface water wave, the rock dropped into the pond is the outside disturbance that starts the wave motion. Once started, the wave continues on its own. We don't have to continue to drop rocks into the pond to keep the ripples moving.

A refined definition of a wave

Let us preface this by saying that it is difficult to come up with a general definition of a wave. For the moment we will content ourselves by looking at the definition of a *material wave*:

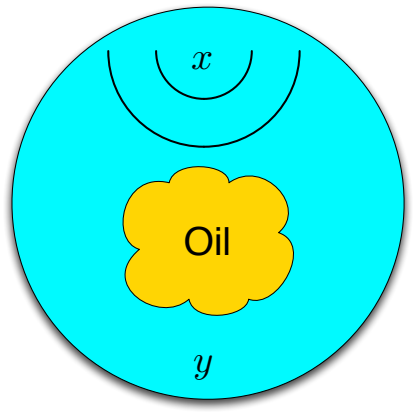
> A *material wave* is a large *movement* of a *disturbance* in the medium from its equilibrium position, whereas the *particles* that make up the medium move very little.

Let's look at each of the italicized words in this definition more closely. A wave is the **movement** of something. But what is it that moves? Think about the example of the expanding ripples in the pond. It is the ripples that move significantly. There is some movement of the individual water molecules, but they merely bob up and down – they do not travel *along* with the wave. The movement of the ripples *along* the surface of the water is what our eyes follow, and those ripples are what we mean when we talk about "the wave". The ripples are a disturbance of the surface of the water. If we

focus on a leaf floating on the pond as the ripples pass, we see the leaf bobs up and down, i.e., oscillates. It does not travel along with the wave. The water surface oscillates as the ripples pass through it, and then stop. But the oscillations are not the wave. Rather it is the ripple (i.e. "disturbance") that moves across the water that we call *the wave*.

Example #1:

Here is an example that will demonstrate the difference between particles of the medium moving versus a disturbance in a medium moving. Take a bowl of water, and tip a small amount of olive oil on the surface, so that you have a situation pictured below:



If you oscillate your hand *gently* at the location "x", do you get waves at the location "y"? Do you end up with oil moving to the location "y"?

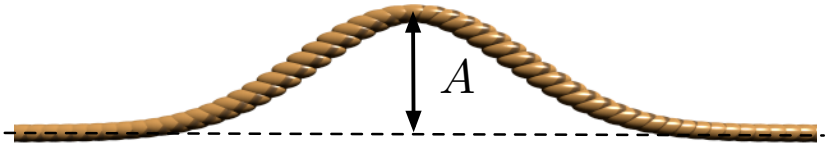
8-1-2 Some properties and characteristics

In most material waves we typically encounter, the "shape" of the disturbance stays the same over short distances of travel. In our example, the ripples look similar as they expand away from the initial disturbance. Over greater distances, however, we notice changes. The ripples seem to die out as the radius of the circle they make increases. In some waves, the shape may remain constant over very long distances, as in low frequency sound waves at large depths of the ocean.

Material waves provide a mechanism for transferring energy over considerable distances, without the transport of the material medium itself.

8-1-2-1 A pulse wave

A pulse is a finite length disturbance. For example, if you shook the end of a rope once you would produce a pulse wave as shown. Another example of a pulse-type wave is the example of a rock thrown into a pond.

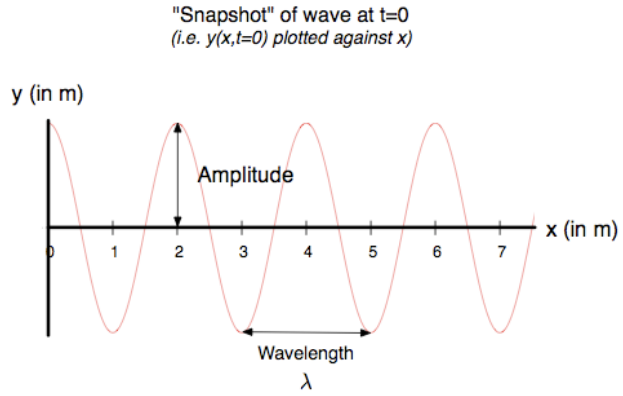


The location of the rope before and after the pulse passes is the *equilibrium position*. The maximum magnitude of the displacement of the pulse from equilibrium is the amplitude, designated here by the letter A. Note that A is always positive.

8-1-2-2 A repetitive wave

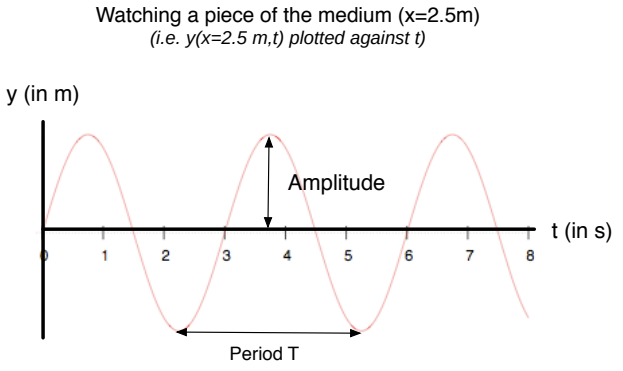
Repetitive waves can have many different shapes. One of the simplest to deal with looks like a sine or cosine function. Such waves are called *harmonic* or *sinusoidal* waves, and are generated by oscillators moving in simple harmonic motion. For example, if you hold one end of a rope and jiggle it up and down in simple harmonic motion, you will generate harmonic waves. If you were to take a picture of the waves, it would look like this:

Here x is the distance along the rope, and y is how far the rope moves sideways as the wave passes through. The curve is the shape of the rope at the instant the picture was taken.



Like the wave pulse, a repeating wave has an *equilibrium position* (the location of the rope when no wave is present) and an amplitude A. In addition a repeating wave has two additional parameters. The first is the *wavelength* λ which tells us how far *along the direction of wave motion* we have to move before the wave looks exactly the same. The second is the period T which tells us how long we have to wait until the wave looks exactly the same. These have no analogue in the pulse-type wave because it does not repeat. A nice way of summarising this is that the wavelength λ tells us how the wave repeats in *space*, while the period T tells us how the wave repeats in *time*.

Notice that the picture of the wave above gives us no information about the period. If we were to paint a red dot on the rope and then plot the position of that dot against time we would find that particular point moves in simple harmonic motion:



Here T is the period of *both* the simple harmonic motion of the red dot of the rope and the wave itself! Just like simple harmonic motion, the period is the reciprocal of the frequency:

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

T is the time between the arrival of adjacent crests of the wave, while f is the number of crests that pass by per second.

This is called a one dimensional wave because is moves only in one direction, here taken as the x-axis. To give examples of higher dimensional waves we refer to the next section.

8-1-2-3 Dimensionality of the wave

We have already mentioned that a wave on a rope is a "one-dimensional wave" because the wave only travels in one direction. If a wave spreads out on a surface then we will define it as a two-dimensional wave. For example a water wave spreads out on the surface of the water so is two-dimensional. A wave that spreads out in all directions is three-dimensional. Examples of three dimensional waves are (typical) sound and light waves.[1]

8-1-2-4 Polarisation

Material waves (and electromagnetic waves) have a characteristic called *polarisation*. The polarisation tells us how the displacements occur in the medium. We are going to break the types of oscillations into two types:

- *Transverse waves*:
  A material wave is transverse if the displacement from equilibrium is perpendicular to the direction the wave is travelling. Note that if we consider a wave travelling to the right of the page then both an oscillation in-and-out of the page *or* toward the top-and-bottom of the page would both be considered transverse. An example of a transverse wave on a spring is shown in figure 8-1.1.

- *Longitudinal waves*:
  A material wave is longitudinal if the displacement from equilibrium is in the same direction that the wave is travelling. An example of a transverse wave on a spring is shown in figure 8-1.2.

[1] Technical aside: these are the definitions that we will use in Physics 7C, although you should be warned that they are not universal. Some books would define a wave that depends only on the distance from the source r as one dimensional, whereas we define it to be three dimensional. However, once there is more than one source then typical light and sound waves are thought of as three dimensional in anyone's definition.

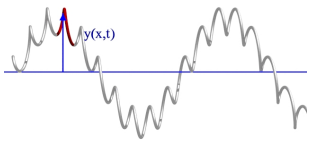


Figure 8-1.1: A transverse wave in a spring. The horizontal line represents the equilibrium position, and the displacement for the coloured coil is explicitly labelled.

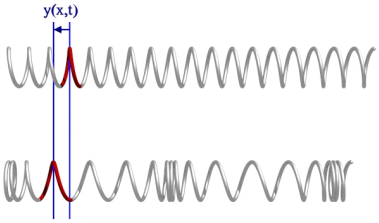


Figure 8-1.2: The spring in its equilibrium position (above) and a transverse wave (below). The displacement from equilibrium $y(x,t)$ can be found be comparing the two pictures. Note that in this case $y(x,t)$ is *not* vertical!

Test yourself:

Which of the following can be seen in figures 8-1.1 and 8-1.2: amplitude, wavelength, period? Which *cannot* be seen on these pictures?

8-1-2-5 Wave speed and "riding the wave"

The wave speed v_{wave} is the speed at which the *disturbance* propagates through the medium. It is *not* the speed of the individual particles making up the medium. One way of thinking about the wave speed is that it is the speed someone who was "riding the wave" on a surfboard would travel.

To a good approximation the wave speed depends only on *properties of the medium*, not on the size of the waves. (For very large waves this approximation breaks down, but we will not be dealing with this in Physics 7C.)

To a good approximation v_{wave} does not depend strongly on the frequency of a repeating wave either. We will simplify the discussion by ignoring any dependence of wave speed on frequency until we discuss rainbows and light.

As an example of how the medium determines the wave speed we can look at a material wave on a stretched medium. Both transverse waves and longitudinal waves are possible on a stretched string or wire. The speed, v_{wave}, of transverse waves on a stretched string depends on the properties of the string that affect its elasticity and its inertial properties. For a string that is thin compared to its length, the relation connecting the wave speed to the string properties is

$$v_{\text{wave}} = \sqrt{\frac{\tau}{\mu}}$$

where τ is the tension in the string and μ is its mass per unit length. Notice that this formula makes some intuitive sense with the picture we discussed earlier. The tension is the (roughly) the force that one piece of string exerts on another – the tighter the string the higher the tension. As we learned that a material wave is a disturbance that propagates by one piece of the medium exerting a force on its neighbours it makes sense that when the tension goes up the wave speed also increases. When the string is particularly heavy, the same force results in less acceleration so it also makes sense that as μ increases the wave speed goes down. The ability to control the wave speed is critical for stringed instruments like the guitar, which is why they have tuning knobs at one end (to control τ) and the strings are of different size (for different μ). We will discuss the guitar in more detail later when we discuss standing waves.

Note that the speed is independent of the time and, if the string is homogeneous it is independent of position as well. Notice also that the wave speed does not depend on how long the string is, nor the amplitude of the wave, nor the frequency of the wave (if it is a repetitive wave) or the shape of the pulse (if it is a pulse-type wave).

8-1-3 Harmonic waves

For the rest of the course we will focus on infinite repeating waves of a specific type: harmonic waves. In a mechanical wave we represent the wave mathematically as

$$y(x,t) - y_0 = A\sin\left(2\pi\frac{t}{T} \pm 2\pi\frac{x}{\lambda} + \phi\right). \tag{8-1.1}$$

The left hand side $(y(x,t) - y_0)$ stands for the displacement of the "particle that would be a distance x along the medium if the rope was in equilibrium at time t". Here y_0 stands for the position of the medium if there was no wave present at all, and $y(x,t)$ is the actual position. On the right hand side we have already seen the parameters A (the amplitude), T (the period), and λ (the wavelength) in §8-1-2-2. The parameter ϕ, called the *fixed phase constant* is new and controls what the wave looks like when $x = 0$ and $t = 0$. Finally $\pm$ means to choose either $+$ or $-$, and controls the direction of the propagation of the wave (i.e. whether the wave moves to the left or the right). To see how this comes about, look at exercise Section 8-1, ex. #2.

Before going too much further, it is worthwhile noting the difference between *variables* and *parameters*.

- The parameters (A, T, λ, ϕ, y_0 and the choice of a "+" or "−" sign) are fixed for any given harmonic wave, and define the wave.
- The wave exists in all space (any x) forever (any t). We pick a specific x and t to ask a question about the displacement of a *specific* piece of the medium at a *specific* time. This distinction makes x and t variables.

We can ask about different locations and times by changing x and t, but the parameters are fixed for the wave.

While we have framed this discussion in terms of material waves because it is the easiest to visualise, we should be aware that the harmonic wave is a much more general concept. It can apply to the variation in pressure (for sound waves):

$$P(x,t) - P_0 = A \sin\left(2\pi\frac{t}{T} \pm 2\pi\frac{x}{\lambda} + \phi\right). \tag{8-1.2}$$

or the variation in electric field in light:

$$\boldsymbol{E}(x,t) = \boldsymbol{A} \sin\left(2\pi\frac{t}{T} \pm 2\pi\frac{x}{\lambda} + \phi\right). \tag{8-1.3}$$

In general we can let $y(x,t)$ stand for any of these physical quantities, not just position. We shall refer to $y(x,t)$ in this general form as the *wave function*. Sometimes harmonic waves are also referred to as *sinusoidal waves* as the wave function is a constant times the sine of an angle.

While waves in the real world do not go on forever (and are created at some initial time) we can still use these harmonic waves as a good approximation, and they offer a considerable simplification. Next we introduce the concept of total phase and show what the "+" and "−" signs do.

8-1-3-1 Total phase Φ

We can rewrite (8-1.1) in the following way:

$$y(x,t) - y_0 = A\sin\Phi(x,t) \tag{8-1.4}$$

where $\Phi(x,t)$ is the *total phase* of the wave and is defined by

$$\Phi(x,t) = 2\pi\frac{t}{T} \pm 2\pi\frac{x}{\lambda} + \phi. \tag{8-1.5}$$

This is useful because all the spatial (i.e. x) and time (t) dependence is in the total phase. Once we know the total phase (from x and t) we can use (8-1.4) to find the displacement from equilibrium.

Because the sine function is periodic with period 2π, changing Φ by 2π, 4π, ... does not change $y(x,t)$. This ambiguity exists partially because the wave keeps repeating so that many places on the wave look exactly the same. Let us try and make our example more concrete: because $\sin\frac{\pi}{2} = 1$ is the maximum of the sine function $\Phi = \pi/2$ labels a peak. Note that $\Phi = 5\pi/2$ also labels a peak, but it labels a *different* peak. When we imagine ourselves "riding the wave" we are really following a *point of constant phase*, as in the next example.

Example #2:

We are going to look at the wave described by

$$y(x,t) = (25\,\text{cm})\sin\left(2\pi\frac{t}{4\,\text{s}} + 2\pi\frac{x}{4\,\text{cm}} + \frac{\pi}{2}\right)$$

a)
One of the peaks of the wave has a total phase $\Phi = \pi/2$. What is the location of this peak when $t = 0$, $t = 1$ s, $t = 2$ s, $t = 3$ s and $t = 4$s?
b)
Is the wave travelling to the left or right?

Solution:

a)

We only need information about the total phase. The only useful information we are given is

$$\Phi(x,t) = 2\pi\frac{t}{4\,\mathrm{s}} + 2\pi\frac{x}{4\,\mathrm{cm}} + \frac{\pi}{2}$$

We are asked to find where (i.e. which x) $\Phi = \pi/2$ is when $t = 0$. We can solve this rather simply:

$$\Phi = \frac{\pi}{2} = 2\pi\frac{0}{4\,\mathrm{s}} + 2\pi\frac{x}{4\,\mathrm{cm}} + \frac{\pi}{2}$$

which is only satisfied for $x = 0$. Therefore the $\Phi = \pi/2$ peak is at $x = 0$ when $t = 0$. Substituting in the other values for t we find

Time t	Peak $\Phi = \pi/2$ located at ...
0 s	0 cm
1 s	−1 cm
2 s	−2 cm
3 s	−3 cm
4 s	−4 cm

b)

We see that a particular peak goes from 0 cm to −4 cm. This peak is moving to the *left*, as are all other parts of the wave.

We can see that the + sign in front of the spatial term is responsible for this. If we remember that by "riding the wave" we are looking at a piece of constant total phase Φ we can that as time increases

$$\underbrace{\Phi}_{\text{not changing}} = \underbrace{2\pi\frac{t}{T}}_{\text{increasing}} + \underbrace{2\pi\frac{x}{\lambda}}_{???} + \underbrace{\phi}_{\text{not changing}} .$$

The only way that the left hand side can remain unchanged is if the second term *decreases* – i.e. the wave travels to the *left*.

Test yourself:

Go through exercise Section 8-1, ex. #2 using the − sign instead.

8-1-3-2 The fixed phase constant ϕ

Note that the phase expression is very similar to the mathematical description we developed for the motion of a particle vibrating in simple harmonic motion. The first term in the argument of the sine function, $2\pi t/T$, is exactly the same as we had for simple harmonic motion. The fixed phase constant ϕ serves exactly the same purpose as there, to give the proper value of y at $t = 0$ and $x = 0$. The only thing new is the term $2\pi x/\lambda$. Note the similarity of this term to the term involving the time. This term involving x and λ gives the change in phase as we look along different values of x. The (total) phase goes through a complete cycle of 2π radians each time x increases or decreases by an amount equal to the wavelength λ. Likewise the (total) phase goes through a complete cycle of 2π radians each time t increases by one period T. This is a reminder that λ controls repetition in *space*, while T controls repetition in *time*.

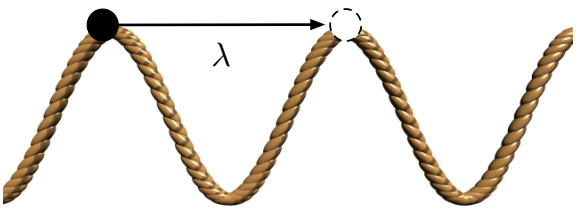
8-1-3-3 Relationship between v_{wave}, λ and f

We have already learned in §8-1-2-5 that the speed of the wave depends on the properties of the medium. When dealing with repeating waves we have three additional parameters: the wavelength λ, the period T and the frequency f. Note that only two of these are independent as we have $f = 1/T$. We will go through two different arguments to show that the frequency, wavelength and wave speed are all related by

$$v_{\text{wave}} = \lambda f. \tag{8-1.6}$$

Distance over time

Our definition of speed from Physics 7B was speed = (distance travelled)/(time taken). Let us look at the wave at a particular time, and focus on a particular peak indicated by the solid dot.



Recall that one period is the shortest amount of time before the wave looks *exactly* the same. If the entire wave moves right one wavelength the peak

indicated by the solid black dot must have moved *one wavelength* to the right to the location of the dashed circle for the wave to look exactly the same. We can now calculate the speed of the peak (which is the same as the speed of the entire wave)

$$\begin{aligned} \text{Distance peak travels} &= \lambda \\ \text{Time taken} = 1 \text{ period} &= T \\ v_{\text{wave}} = \frac{\text{Distance travelled}}{\text{time taken}} &= \frac{\lambda}{T} \end{aligned}$$

Using the fact that $1/T$ is another way of writing the frequency f, we can write this formula in a more familiar form:

$$v_{\text{wave}} = \lambda f$$

If the wave is moving to the left this works also, the only modification to the argument is that the "dashed circle" would sit one peak to the left of the original peak shown. We would still get one wavelength travelled in one period, so the speed is still $v_{\text{wave}} = \lambda/T$.

Following the total phase Φ

A superficially different way of finding the wave speed is to follow a piece of the wave, that is look at a piece of the wave with a constant total phase Φ. We did this already in example Section 8-1, ex. #2. Let us pick a phase Φ and look at it at two different times t_1 (where it is at x_1) and t_2 (where that piece of the wave is at x_2). This gives us the relationships

$$\begin{aligned} \Phi &= 2\pi\frac{t_1}{T} \pm 2\pi\frac{x_1}{\lambda} + \phi \\ \Phi &= 2\pi\frac{t_2}{T} \pm 2\pi\frac{x_2}{\lambda} + \phi \end{aligned}$$

Now we can subtract these equations from one another:

$$2\pi\frac{t_2 - t_1}{T} \pm 2\pi\frac{x_2 - x_1}{\lambda} = 0$$

or written another way

$$2\pi\frac{\Delta t}{T} \pm 2\pi\frac{\Delta x}{\lambda} = 0$$

where we have used the Δ for "final minus initial". Cancelling the 2π and rearranging we have

$$\frac{\Delta x}{\Delta t} = \mp\frac{\lambda}{T}$$

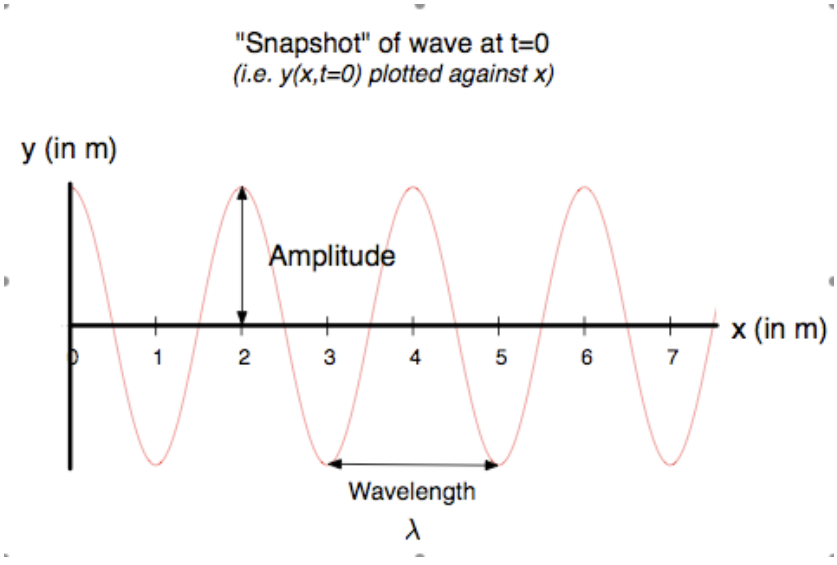
But this expression tells us how far (Δx) the "disturbance" with phase Φ moved in a time Δt. This is exactly what we mean by velocity! Taking the absolute value of this gives us the wave *speed*:

$$v_{\text{wave}} = \left|\frac{\text{distance travelled}}{\text{time taken}}\right| = \left|\frac{\Delta x}{\Delta t}\right| = \frac{\lambda}{T} = \lambda f$$

8-1-4 Graphical representations of waves

Now one of the tricky things about the solution to the wave equation expressed in equation (8-1.1) is that it is a function of both space (the distance along the x axis) and a function of time (the value of the time variable, t). One way to make visualise the equation is to think of either x or t as being fixed at some value, so that y depends one only one variable. This is exactly what we must do in order to graph the function $y(x,t)$ in a simple 2D graph. We will describe these two graphs for the same example wave below.

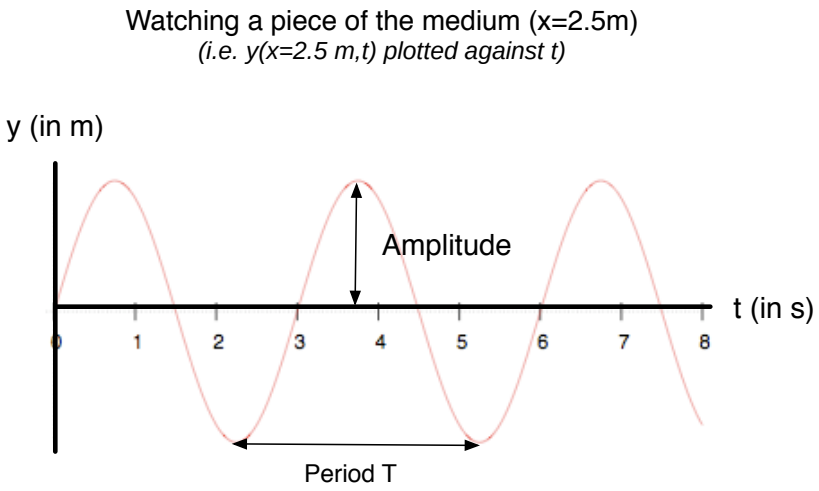
- **Holding t constant: "Displacement versus position"**



This graph shows the displacement of the entire wave at a particular time, thus the name "snapshot". For a longitudinal wave (e.g. figure 8-1.2) this graph does not really look like the wave, but for a transverse

wave (e.g. figure 8-1.1) the name "snapshot" is apt. This graph is correct even for longitudinal waves, even though it does not correspond directly to a visualisation.

- **Holding x constant: "Displacement versus time"**



This graph focuses on a particular piece of a medium and plotting its motion against time. One way of visualising this is tying a ribbon to a piece of the medium, and then plotting how the ribbon moves in time. This graph reminds us that this particular piece of the medium is undergoing simple harmonic motion.

To convey all of the information contained in equation (8-1.1), requires both graphs. Alternatively, almost all of the same information can be conveyed using two displacement versus position graphs for two different times, or two displacement versus time graphs with each for a distinct position along the hose.[2]

Test yourself:

Using the two graphs above, can you find the equation for the wave? You should be able to get *numerical* values for all the variables except one – which

[2]The difference is that if we have two displacement versus position graphs (for example) at different times, we don't know for sure which peak in one graph has the same phase as a particular peak in the other graph. This leads to some ambiguity in what the period actually is. The case where there is no ambiguity is when we have a displacement versus time graph *and* a displacement versus position graph.

one can't you determine? Hint: Which way is the wave travelling (left or right)? You will need this to determine the + or − sign. Which parameters do you get from which graph? Which parameters require *both* graphs?

8-1-5 Other types of waves

So far we have concentrated on waves that are movements of a medium. In Physics 7B we discussed how an object close to equilibrium underwent simple harmonic motion if the displacement was small (and the technical condition that the equilibrium was *stable*). At the beginning of this discussion we illustrated how a material wave arose: as pieces of the medium were displaced they pulled or pushed on the next piece of the medium and even though we only disturb one piece of the medium other pieces can be affected. This travelling disturbance was what we referred to as the wave. Because for small oscillations each part of the medium undergoes simple harmonic motion (as we discussed in Physics 7B) it does not seem all that surprising that all these different media can support waves.

You have probably heard the term "sound waves". That is because sound, like the material waves we have already discussed, is literally a wave in the sense that we have already described. In this case the medium is the material the sounds travels through. Often this is the air, but sound can also travel through liquids and solids. As we discussed in Physics 7A the bonds between particles in the air are virtually non-existent, so it is difficult to think of the air molecules as having an equilibrium position. Instead, as we discuss shortly, we tend to think about sound as a pressure wave. While this is slightly different in form from the material waves we have already discussed, we can apply almost all of the same techniques that we have already learned to sound waves. Fundamentally, however, sound waves are just material waves in a medium and so we may not be surprised that the same techniques work.

Later on in the course we will also discuss light and quantum particles. In each of these cases we have "waves" that have no medium, and so are *not* propagated along by forces pulling and pushing a material. Why do we call them waves then? The answer is that these light and quantum waves exhibit much of the same behaviour as the material waves we are currently discussing. In this unit we will show how to combine material waves (interference), talk about how material waves travel (diffraction and geometric optics). The light

and quantum waves have these properties too! We take the view that we will use material waves to build up our intuition, and then define a *general wave* as anything that has similar behaviour to the material waves we have come to know.

A wave model of sound

We can choose to model sound in two different ways. The first model is that the individual molecules that make up the air are vibrating back in forth in the direction of propagation. That is, sound waves are *longitudinal*. We know that the position of an air molecule is described by (8-1.1), but now $y(x,t)$ does not tell us the height of the particle (as it would for a *transverse wave*) but tells us how the displacement of the particle from its equilibrium position.[3] This is illustrated in figure 8-1.2. We have already learned while discussing material waves how to model such waves, and in this sense sound waves are simply longitudinal material waves.

There is a separate way of thinking about sound waves. Instead of focussing on the position of the air molecules (which is difficult to measure experimentally) we can instead measure either the *density* or *pressure* of the air. Recall from what we learned in Physics 7A that in a room in equilibrium the air molecules are in random motion, moving around rapidly, but there are well defined averages for density and pressure. Sound waves in air involves the oscillation of the average value of particle density (and resulting pressure) over distance scales much larger than the mean distance between particles. Thus we can also choose to describe a sound wave in terms of either the pressure or density of the air. Choosing pressure, we can describe sound by

$$P(x,t) = A\sin\left(\frac{2\pi}{T}t \pm \frac{2\pi}{\lambda}x + \phi\right) + P_{\text{atm}} \tag{8-1.7}$$

where $P(x,t)$ is the absolute pressure of the air at a given position x along the tube, and at a time t. P_{atm} is the equilibrium pressure (i.e., atmospheric pressure), and A is the amplitude of the pressure fluctuation (gauge pressure) from equilibrium.

[3]This contradicts the statement we made on the previous page that the atoms are zooming around and don't have an equilibrium position. The previous page *is* correct, but the motion of the molecules is random and does not contribute overall. We can get a good description of sound by treating the air molecules as vibrating back and forward, but the pressure description described below is a better "fit" to reality.

Note the similarity between this equation and the equation (8-1.1) for $y(x,t)$. We can use the same techniques to plot the pressure against time at constant position, or pressure against position at constant time.

8-1-6 Summary

A detailed summary of the properties of waves and the types of waves they apply to are summarised in table 8-1.1. Notice that all of these properties apply to harmonic waves, except polarisation for some special types of waves. The main concepts from this chapter, in addition to the terms defined above, are:

1. A material wave is a propagating disturbance in a material, while the atoms that make up that material do not travel very far.

2. Waves describe a large range of phenomena such as ripples in a medium, pressure fluctuations in sound or even fluctuations that describe light.

3. The idea of a wave function $y(x,t)$ that describes displacement, or pressure, ...

4. Harmonic waves which have a wave function given by

$$\Delta y = y(x,t) - y_0 = A \sin \Phi(x,t)$$

   where y_0 is the equilibrium value of $y(x,t)$ and Φ is the total phase.

5. The wave function and the total phase are functions of space *and* time; knowing only one is not good enough.

6. Two common representations of waves: $y(x, t = \text{const})$ vs x **or** $y(x = \text{const}, t)$ vs t and what these graphs physically correspond to.

7. The wave speed v_{wave} is set by the medium.

8. The frequency f is set by the source.

9. The wavelength depends on both the frequency and velocity: $\lambda = v_{\text{wave}}/f$.

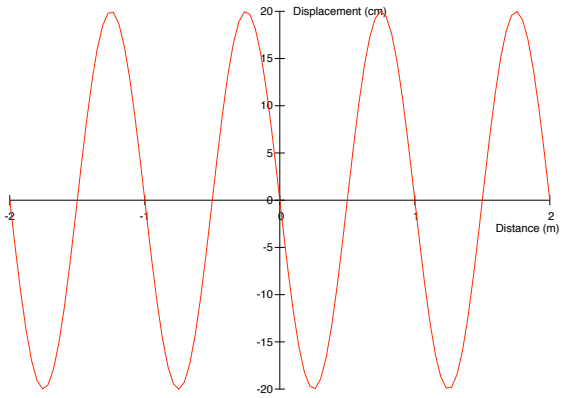
Almost all of physics 7C builds upon the main ideas presented above, so make sure you have a solid grasp of them!

Property		Applies to	Description
Amplitude	A	All waves	The maximum displacement from equilibrium.
Speed	v_{wave}	All waves	The speed at which a disturbance moves.
Dimensionality		All waves	- 1-D: wave travels along a line (e.g. waves on a rope). - 2-D: wave spreads out over a surface (e.g. water waves). - 3-D: wave spreads out over space (e.g. sound waves).
Direction		All waves	Direction the wave travels. Determined by the $+$ or $-$ sign for 1-D waves.
Polarisation		Most waves	- Longitudinal: vibrations in the direction of wave propagation. - Transverse: vibrations perpendicular to the direction of wave propagation.
Period	T	Repetitive	Time taken for wave to look exactly the same.
Frequency	f	Repetitive	$f = 1/T$
Wavelength	λ	Repetitive	Shortest distance along the wave before the wave looks exactly the same.
Fixed phase	ϕ	Harmonic	Sets conditions when $t = 0$ and $x = 0$.
Total phase	Φ	Harmonic	

Table 8-1.1: A list of wave properties and the types of waves they are associated with. Notice that harmonic waves are a special case of repetitive waves.

8-1-7 Exercises

1. The following questions relate to the wave shown below:



(a) A "snapshot" of a wave at $t = 0$ is taken and shown above. What

is the phase constant if the wave is travelling to the right? What is the phase constant if the wave is travelling to the left?

(b) Using the same snapshot, what is the amplitude if the wave is travelling to the right? What is it if the wave is travelling to the left?

(c) What can you tell about the wavelength by looking at this graph? What about the period? Do these depend on whether the wave is travelling to the right or left?

2. Two ropes of different densities are tied together to form a long rope. We are going to create a continuous wave by moving one end of the rope in simple harmonic motion. For each of the statements, is the answer true or false?

(a) The wave speed v_{wave} *must* be the same in both ropes.

(b) The frequency f *must* be the same in both ropes.

(c) The wavelength λ *must* be the same in both ropes.

(d) The amplitude A *must* be the same in both ropes. (Tricker, not directly covered in notes. Think of an extreme limit of a string rope and a steel cable.)

3. Why is a mass on a spring *not* a wave? Which parameters does it have in common with a harmonic wave? Which make sense for a harmonic wave but do not make sense for a mass on a spring?

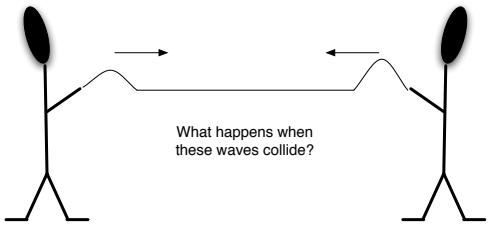
Unit 8:

8-2: Superposition and Interference

8-2-1 Overview

So far in the course we have discussed how a source could create a wave pulse, a repeating wave or a harmonic wave. By knowing the motion of the source, we have seen that the disturbance keeps its shape and propagates with a speed v_{wave}.[1] These discussions all assumed that the medium was flat before the wave propagated – what if there was *another wave* already in the medium? What happens when the two waves collide?

An example of how this could occur is if you and a friend both hold a rope. If you wiggle your end, the wave you make will propagate toward your friend. If your friend propagates her end, her wave will propagate toward you. What happens when your waves collide?



A different example is dropping two stones in a river. Eventually the ripples will overlap; how can we calculate the displacement from equilibrium?

[1] Strictly this is only true if the speed of the waves does not depend on frequency. We emphasised earlier that v_{wave} does not depend on f to a good approximation. We will have to deal a situation where v_{wave} depends on f when we discuss rainbows later.

The solution to this problem is to do the following: consider each displacement from equilibrium separately. For each piece of rope *add* (as a vector) the displacement for each wave. This "combined displacement" will be the total displacement for that piece of rope. This procedure is valid provided the amplitude of the wave is small. For example, if your friend's wave would have caused a particular piece of the rope to rise 2 cm, and your wave caused the same piece of rope to rise 1 cm, the actual amount that piece of rope will rise is 3 cm. The idea of adding the individual effects of waves to get the total effect is called *superposition*. With the exception of some information on phase changes and reflection in §8-2-5-1, the idea of superposition is the only new piece of information in this entire section. The rest of this section is devoted to the implications of superposition.

8-2-2 The idea of superposition

In the overview we gave the general idea of superposition, in this section we will simply be making that idea more precise and introducing some of the language used discussing superposition. Superposition is the idea of adding the effects of two (or more) waves together at *the same location at the same time*. This gives us the total effect of the two waves. It makes *no sense* to add what happens due to a wave at one location to what happens to a wave at another location.

For material waves, we can replace the word "effect" with the word displacement, although the principle of superposition works for non-material waves (such as electromagnetic waves, the pressure interpretation of sound and matter waves). For the time being, let us concentrate on material waves. We express superposition mathematically as follows:

$$\Delta y_{\text{tot}}(x,t) = \Delta y_1(x,t) + \Delta y_2(x,t)$$

where Δy_1 and Δy_2 are the displacement from equilibrium for wave 1 and wave 2 *only*. The actual displacement of the medium is described by Δy_{tot}. We illustrate this procedure in figure 8-2.1.

Conventions on space and time

Some of our conventions are useful, but a little confusing at first. We have emphasised already that superposition is combining two or more waves acting at the same location at the same time. But so far we have dealt only with single source systems, and we have always chosen the origin of our coordinates

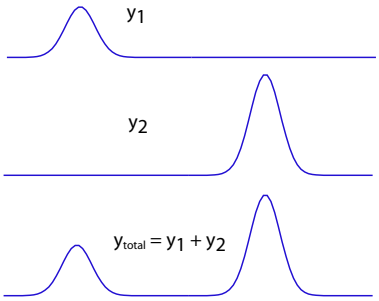


Figure 8-2.1: Illustrating superposition. We add the waves represented by Δy_1 and Δy_2 to get the superposed wave shown in the bottom picture.

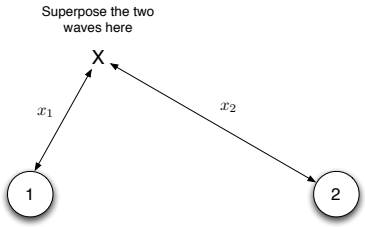


Figure 8-2.2: To figure out the wave at the point labelled "x" we add the effect of the wave from source 1 to the effect of the wave from source 2 *at the location "x" at the same time.*

to be the location of the source. What do we do if we wish to combine the effects of the two sources (indicated by circles) at the location of the "x" in figure 8-2.2? To keep as close as possible to the work we have already done on waves, we adopt the following conventions:

- We use a universal clock t. As we are combining the effect of the two waves at the *same* time, we should use the *same* value t in Δy_1 and Δy_2.
- We use a different origin for each source. Even though we are combining the waves at the same location, we have *two* distances x_1 and x_2. Here x_1 is the distance between source 1 and where we wish to combine the waves; an analogous definition holds for x_2. We would use x_1 for calculating Δy_1 and x_2 for calculating Δy_2 even though we are interested in the same point.

When we are using a sinusoidal wave we also need a convention for ϕ_1 and ϕ_2, the phase constants. The convention we use here is that ϕ_1 determines

what the source is doing (i.e. $x_1 = 0$) at $t = 0$. Analogous comments hold for ϕ_2.

The reason that we use these particular conventions, rather than just picking one origin, is that it allows us to keep the formulas

$$\Delta y_1 = A_1 \sin\left(\frac{2\pi t}{T_1} \pm_1 \frac{2\pi x_1}{\lambda_1} + \phi_1\right)$$
$$\Delta y_2 = A_2 \sin\left(\frac{2\pi t}{T_2} \pm_2 \frac{2\pi x_2}{\lambda_2} + \phi_2\right)$$

that we are so used to when we specialise to sinusoidal waves in §8-2-3. (Here $\pm_1$ and $\pm_2$ refer to the direction of propagation of wave 1 and 2 respectively, and are independent.)

There is one more convention that is worth noting: we treat x_i as a *positive* distance from the source. For a wave that travels outward (this is almost always the case) we would use the $-$ sign. This is because the peak of a wave (for example) gets further away from the source as time increases. We would only use the $+$ sign when waves were travelling inward. The following illustration may help with the change from what we did in the first section:

The vocabulary of interference and superposition

While the idea of superposition is fairly straightforward, there is a lot of associated vocabulary that comes with it. Intuitively we can see that if two waves are oscillating up and down together that the resulting oscillations from combining the two waves will be larger. This is called *constructive interference*. On the other hand, if one wave is going up while the other is going down then the two waves are cancelling each other out. This is known as *destructive interference*. If the waves are the same amplitude, then these waves will cancel each other out completely! It is also possible that waves are neither completely in step or out of step, which we refer to as *partial interference*. Partial interference is not very descriptive – we can have partial interference that is either *almost* constructive or *almost* destructive.

The experimental status of superposition*

Because simply adding the waves together is the most obvious thing to do, it is worth pausing and considering if it is the *only* way we could have combined

the effects of two waves. The answer is "no" – we could have combined the waves in much more complicated ways. For *very* large water waves or sound waves we cannot simply use the principle of superposition presented here. Shockwaves, such as the ones produced by explosions or sonic booms, are examples of waves for which the principle of superposition simply does not work.

We are lucky that for "small" waves the principle of superposition is adequate. But it should be appreciated that this is an experimental result, and not one that can be derived from purely logical thought.

8-2-3 Specialising superposition to harmonic waves

While the idea of superposition is relatively straightforward, actually adding the displacements of the waves at every point for all time is a lot of tedious work. We are now going to specialise to the case where we are dealing with infinite sinusoidal waves. Instead of keeping track of both the wavefunctions Δy_1 and Δy_2 this means that we only have to look at the difference in total phase $\Delta\Phi$.

For example if we know that at a particular location the peaks of both waves arrive simultaneously, and the troughs of both waves are occurring simultaneously then we would say the waves are in phase. Our obvious guess would be that $\Delta\Phi \equiv \Phi_2 - \Phi_1 = 0$ because the peaks and the troughs are arriving together. However we know that if the total phase changes by 2π, 4π, 6π, ... that the wave looks exactly the same – a fact that can be traced back to the fact that the sine function repeats every 2π. If we have constructive interference all we know is that $\Delta\Phi = 0$ **or** 2π **or** -2π **or** 4π **or** ...

To see how $\Delta\Phi$ tells us about the type of interference it helps to recall two important points about the sine function:

- The sine function is periodic so that $\sin(\Phi) = \sin(\Phi + 2\pi n)$ where n is any integer. This is the point being made in the paragraph above.
- $\sin(\Phi + \pi) = -\sin(\Phi)$, or that half a period along the sine function changes sign (but has the same magnitude). By mathematical induction the same result holds if we replace π by 3π, 5π, 7π, ...

The total displacement we can find by looking at

$$\Delta y_{\text{total}}(x,t) = A_1 \sin(\Phi_1) + A_2 \sin(\Phi_2)$$

When the sines are the same (i.e. $\Delta\Phi = (\text{even}) \times \pi$) we have constructive interference, when the sines have opposite signs (i.e. $\Delta\Phi = (\text{odd}) \times \pi$) we have destructive interference. Anything else is partial. This table summarises the types of interference condition you can get:

Interference type	$\Delta\Phi$
Constructive	(even) $\times\pi$
Destructive	(odd) $\times\pi$
Partial	(non-integer) $\times\pi$

Recall that the total phase $\Phi(x,t)$ for each wave depends on both x and t, so $\Delta\Phi$ can depend on both x and t. Strictly speaking we should not talk about whether two waves have constructive, destructive or partial interference, but rather *if two waves at a specific location, at a specific time*, have constructive, destructive or partial interference.

To keep track of all the terms that can contribute to the change in phase, we introduce the *phase chart*. The phase chart contains no more information than the three equations

$$\begin{aligned}
\Phi_1 &= 2\pi\frac{t}{T_1} \pm 2\pi\frac{x_1}{\lambda_1} + \phi_1 \\
\Phi_2 &= 2\pi\frac{t}{T_2} \pm 2\pi\frac{x_2}{\lambda_2} + \phi_2 \\
\Delta\Phi &= \Phi_2 - \Phi_1,
\end{aligned}$$

but is meant to remind you to think about each term. The phase chart is shown below:

	$2\pi\frac{t}{T}$	$\pm 2\pi\frac{x}{\lambda}$	ϕ	Φ
Wave 1				
Wave 2				
Change				$\Delta\Phi$

← add first three columns to fill in last column

↑ ↑ ↑ ↑

Find bottom row by subtracting 2nd row from 1st

It is the lower-right hand corner of this chart in the box marked $\Delta\Phi$ that determines if the interference is constructive, destructive or partial.

To build our intuition we are going to look at simplified examples and study the effect of each piece. In the next three subsections we will study:

- **The effect of spatial difference**: By keeping the sources creating waves *in phase* and at the *same frequency*, we can study what the effect is of moving the detector around.
- **The effect of spatial difference and changing the sources**: By keeping the same frequency and amplitude, but now allowing the sources to create waves that are not in phase we can study how out of sync sources affect interference.
- **Beats**: Looking at the interference of two waves with *different* frequencies.

There are a couple of other general comments to make. The first is that because we are combining waves in the same place, the waves must be in the same medium. Therefore the two waves have the same wave speed v_{wave}. Because they have the same wave speed and same frequency, they must have the same wavelength $\lambda = v_{\text{wave}}/f$. The periods of the two waves $T = 1/f$ must also be the same. The quantities which may be different are the distance from the source to the detector x, and the phase constant ϕ. By changing either of these quantities we can have either constructive or destructive interference.

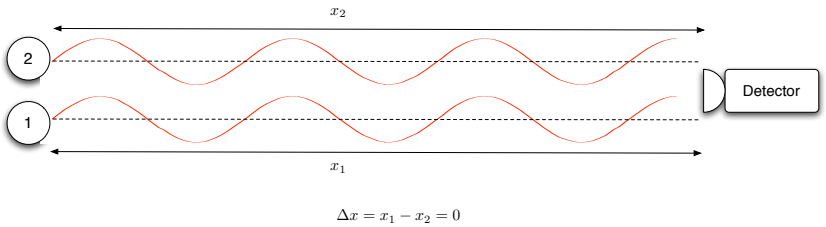
8-2-3-1 Path length difference

For this part of the notes we will assume that the two waves are at the *same* frequency and have the *same* amplitude. We are also going to assume that the two sources are *in phase* with one another. The most important assumption is that the frequencies are the same, and we should discuss the consequences of this assumption before doing anything else. By having the same frequency we know that the waves both have the same period ($T = 1/f$). Thus, if the waves started oscillating "in phase" at a particular location they will *always* be oscillating in phase because they are both oscillating at the same rate. Likewise if they started oscillating "out of phase" they will remain out of phase. One way of summarising this is that assuming the frequencies are the same means the type of interference depends on *where* you are, but unlike the completely general case does not depend on *when* you ask about the type of interference. To see this from the phase chart we notice that the two terms containing time are exactly the same, so the time part does not contribute anything to the change in total phase:

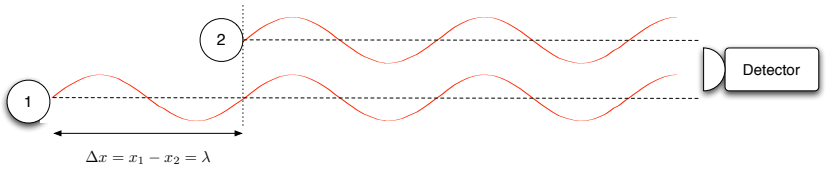
	$2\pi\frac{t}{T}$	$\pm 2\pi\frac{x}{\lambda}$	ϕ	Φ
Wave 1	$2\pi\frac{t}{T}$			
Wave 2	$2\pi\frac{t}{T}$			
Change	0			$\Delta\Phi$

As both our waves are travelling in the same medium we know that $v_{\text{wave}} = f\lambda$ is the same. Because the frequency is the same and the speed v_{wave} is the same both sources must have the same wavelength λ. As we can see, assuming that the sources produce waves of the same frequency leads to a large simplification.

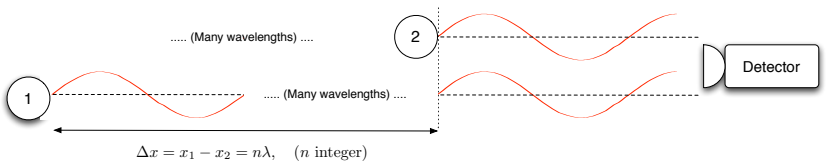
Let us start with two sources that are creating waves *in phase* with one another, and located the same distance from the detector. A picture of the situation may look like the one below:



By adding these waves together we see that the total wave will have twice the amplitude at the detector – this is constructive interference. The waves at the detector look identical if we shift one of the sources one wavelength closer to the detector; this is because after one wavelength the wave looks exactly the same.



Shifting the source by additional whole numbers of wavelengths still leads to constructive interference, as the waves still look identical after shifting:

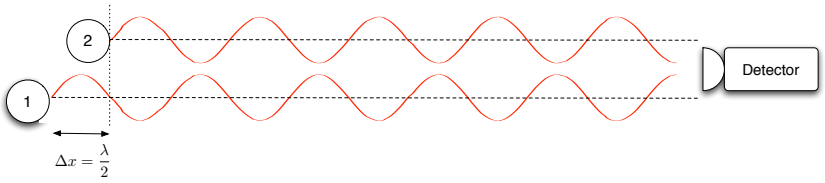


The idea that shifting by a whole wavelength not changing the shape of the wave is intuitive, but how does it relate to the change in phase? Assuming both waves are propagating outward (so we may use the $-$ sign) we have

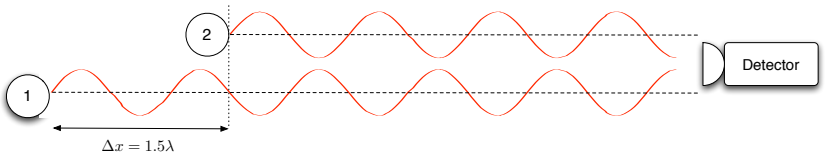
$$\begin{aligned}\Delta x &= x_1 - x_2 = n\lambda, \quad (n \text{ an integer, number of wavelengths shifted}) \\ \Delta\Phi &= -2\pi\frac{\Delta x}{\lambda} + \Delta\phi \\ &= -2\pi\frac{n\lambda}{\lambda} + 0 \\ &= -(2\pi)n.\end{aligned}$$

In the third line we used the fact that the sources were in phase (meaning that they were creating peaks together, and creating troughs together) so $\Delta\phi = 0$. The quantity Δx tells us how much further wave 1 had to travel to reach the detector than wave 2, and is referred to as the *path length difference*.

By shifting one of the sources half a wavelength closer to the detector, we ensure that every peak in wave 1 coincides with a trough in wave 2. In fact, the waves added together completely cancel each other out and we have destructive interference:



Changing the separation by a wavelength (i.e. so the *total* separation is one and a half wavelengths) does not change what the waves look like at the detector, so we still have destructive interference.



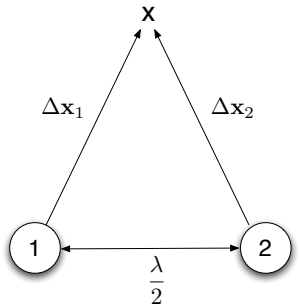
In fact, it is not difficult to see that having $(n + 1/2)\lambda$ as the path length difference will lead to destructive interference. To see this is consistent with

the phase difference picture we calculate $\Delta\Phi$:

$$\begin{aligned}\Delta\Phi &= -2\pi\frac{\Delta x}{\lambda} + \Delta\phi \\ &= -2\pi\frac{\left(n+\frac{1}{2}\right)\lambda}{\lambda} + 0 \\ &= -2\pi\left(n+\frac{1}{2}\right) = -2\pi n + \pi = (\text{odd})\pi\end{aligned}$$

Note that $\Delta\phi = 0$ still, as the waves are still creating peaks (or troughs) at the *same time* as one another. By having different separations we can create the waves in phase (we say the sources are in phase) but have destructive interference when the two waves come together.

It is important to distinguish the separation of the sources and the path length difference. In all of the above examples, these are the same. Consider two sources separated by half a wavelength, but place the detector equal distances from both sources:



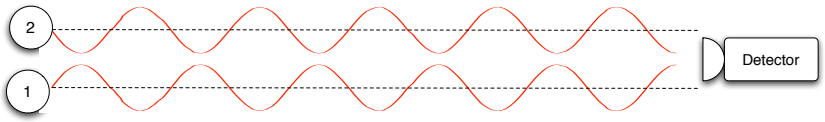
Using vector subtraction:$|\Delta\mathbf{x}_1 - \Delta\mathbf{x}_2| = \lambda/2$
Path-length difference: $|\Delta\mathbf{x}_1| - |\Delta\mathbf{x}_2| = 0$

Now even though the sources are separated by $\lambda/2$, the wave from each source must travel exactly the same distance to get to the detector. Therefore the *path length difference* is zero – peaks created at the same time will arrive at the same time and we still have *constructive* interference. Even though we *can* think of Δx as a vector quantity as we did in Physics 7B we don't want to – the only thing of interest is how far the waves travel from there source. From here on we shall dispense with the absolute value signs.

8-2-3-2 Constant phase differences

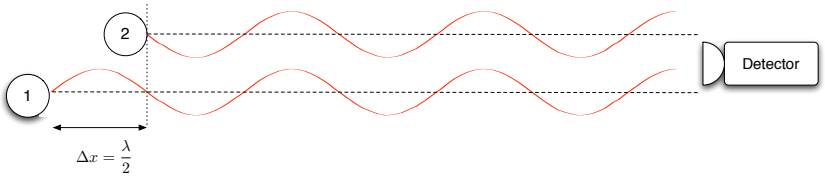
Another way of changing the total phase is ensuring that the two sources are not creating peaks together. This is done by manipulating ϕ_1 and ϕ_2, the

phase constant. To keep things simple we are again going to assume that the frequencies (and hence the wavelengths) of the waves produced by the two sources are the same. If we start with two sources in the same location, but make one source create a trough while the other source creates a peak we have destructive interference at the location of the detector:



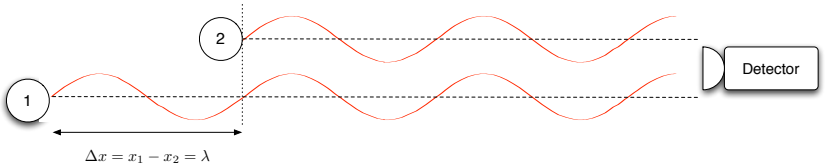
We call these two sources "out of phase". $\Delta\phi = \pi$.

We can combine the two effects shown above. Let us have two sources "completely out of phase" ($\Delta\phi = \pi$) so that source one is creating a peak while source 2 is creating a trough. In addition, we will place source 2 half a wavelength ahead of source one as shown:



8-2-3-3 Using phase charts

Let us see how we can reproduce some of the results we had earlier. Let us look at the case were we had the two sources in phase ($\phi_1 = \phi_2 \equiv \phi$), but the sources were separated by one wavelength:



For the two phases we have

$$\Phi_1 = 2\pi\frac{t}{T} - 2\pi\frac{x_1}{\lambda} + \phi$$
$$\Phi_2 = 2\pi\frac{t}{T} - 2\pi\frac{x_2}{\lambda} + \phi$$

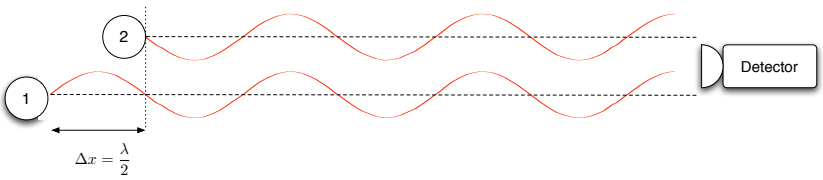
Because these waves have the same frequency they have the same period ($T = 1/f$), and because they are travelling in the same medium they have

the same v_{wave} and hence the same wavelength ($\lambda = v_{\text{wave}}/f$). The difference in phase is

$$\begin{aligned}\Delta\Phi = \Phi_1 - \Phi_2 &= 2\pi t\left(\frac{1}{T} - \frac{1}{T}\right) - \frac{2\pi}{\lambda}(x_1 - x_2) + (\phi - \phi) \\ &= -\frac{2\pi}{\lambda}\Delta x\end{aligned}$$

By looking at the figure above, we see that $\Delta x = \lambda$, so that $\Delta\Phi = -2\pi$ giving us constructive interference. Of course, the easy way of doing this problem would be to look at the waves at the detector – they are obviously in phase so the interference must be constructive!

As a second example, let us consider the case where the two waves were out of phase and separated by half a wavelength as shown:



This time we have $\phi_1 - \phi_2 = \pi$, and $x_1 - x_2 = \lambda/2$. The change in phase is

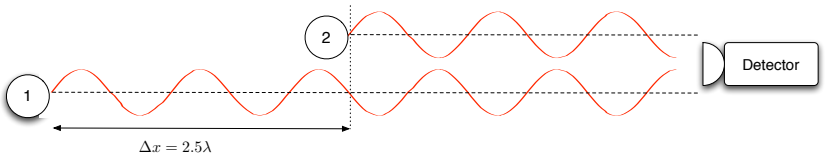
$$\begin{aligned}\Delta\Phi &= 2\pi t\left(\frac{1}{T} - \frac{1}{T}\right) - \frac{2\pi}{\lambda}(x_1 - x_2) + (\phi_1 - \phi_2) \\ &= 2\pi t(0) - \frac{2\pi}{\lambda}\frac{\lambda}{2} + \pi \\ &= 0 - \pi + \pi = 0\end{aligned}$$

which gives us constructive interference again.

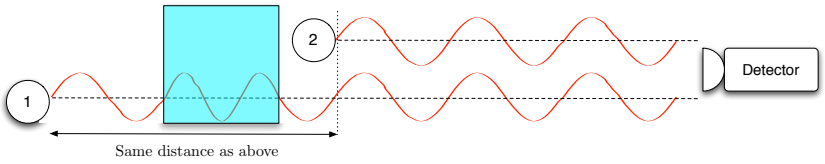
In order make sure that you include every term in the phase difference, you may find a *phase chart* useful.

8-2-3-4 When phase charts fail*

The phase chart is a slight oversimplification of what occurs in the real world. For example, consider two sources are in phase, but separated by 2.5 wavelengths:



We see that we get destructive interference just as we expect. Keeping the same separation, but inserting another medium (shown in blue) leads to a shorter wavelength in that medium



Now the waves are back in step. Notice that at the detector the wavelengths are identical: $\lambda_1 = \lambda_2$. The assumption that we are making in using the phase chart is that the two waves have the same wavelength between the source and the detector. We can get rid of this assumption by carefully keeping track of the phase from one medium to another, i.e. using the equations directly and disposing of the phase chart.

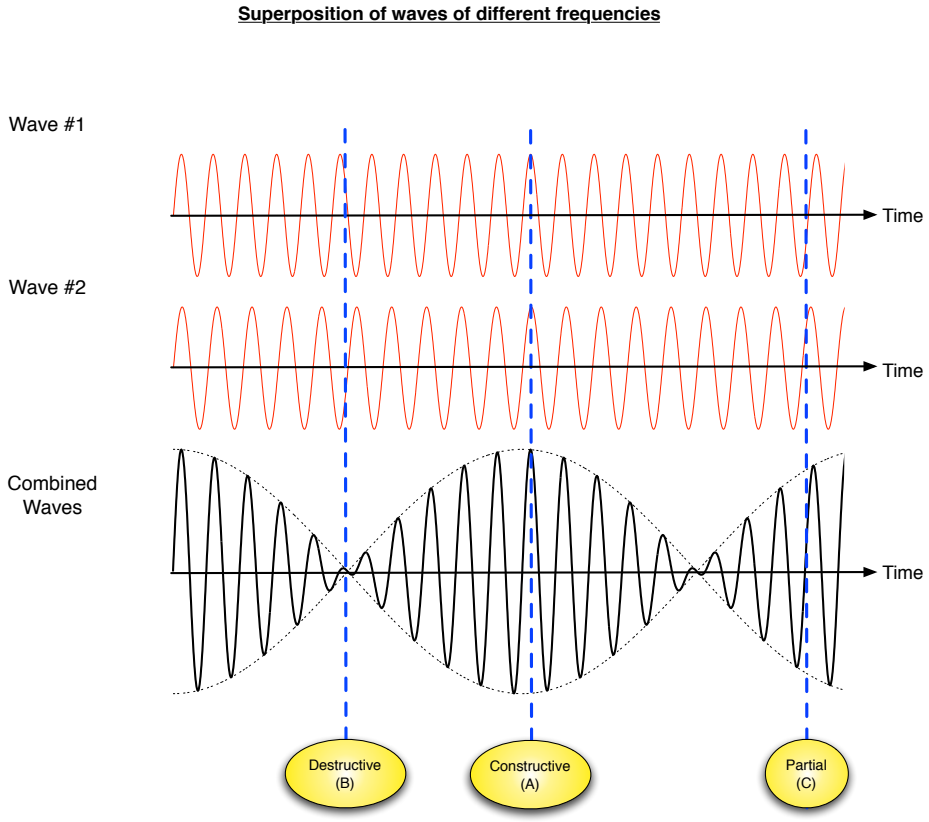
Is this objection relevant to any real world examples? Yes! The subject of *thin-film* interference is based around light interfering, where one ray goes through two mediums and the other ray only goes through one. Thin film interference is responsible for the pretty colours we see on soap bubbles and in puddles on the street where small amounts of oil sit on the surface. Thin film interference is also responsible for the different colours that are seen reflecting in the surface of pearls (the layers are calcium carbonate and water). Thin film interference finds important applications in photography as well.

8-2-4 Beats

So far we have dealt with waves of the same frequency, which leads to a significant simplification: if the waves are in phase, then because they are oscillating at the same rate they stay in phase. We then get constructive interference all the time. Similar comments hold if we have the two waves producing destructive interference. The way this arises mathematically is that the "time term" $2\pi t/T$ is the same for both waves and so contributes nothing to the *change in total phase*, so the type of interference we get is independent of time.

What if the frequencies and periods are different? Consider an example where one wave has a period of 5 seconds, and the other has a period of 6 seconds, and we measure the displacement of the medium at a particular location against time. After 5 seconds, one wave has completed a cycle, but the other has not. If they started in phase (giving constructive interference) they are now not quite in phase so the interference is only partial. After 15 seconds, one of the waves has completed 3 cycles ($5\,\mathrm{s} \times 3 = 15\,\mathrm{s}$) but the other wave has only completed 2.5 cycles ($6\,\mathrm{s} \times 2.5 = 15\,\mathrm{s}$). If the waves were constructive initially, they are now destructive! The fact that the type of interference leads to constructive and then destructive interference is termed *beats*. This comes from the fact that in sound the amplitude is going from large at constructive interference (and hence loud) to small at destructive interference (and hence soft).

Now that we have described what we expect to see, let us actually plot out two waves and add them together. Recall we are taking two waves that reach the same location, then plotting Δy against time.



The dashed vertical lines emphasise that we are adding together the displacement caused by wave 1 to the displacement caused by wave 2 *at the same time*. Remember that throughout this section, this is the only idea that we are using!

Let us make a few comments about this graph. We can see that the waves go gradually from in step to out of step and back again (the vertical dashed lines at points A, B and C can serve as a guide to your eyes). The dashed line around the outside of the two curves roughly tells us how far in step or out of step the two waves are. If the two wave frequencies are f_1 and f_2 then the waves go from constructive to destructive and back to constructive $|f_1 - f_2|$ times per second. We call this the *beat frequency*:

$$f_{\text{beat}} = |f_1 - f_2|, \quad \# \text{ constructive to constructive cycles per second}$$

There is also a *beat period* $T_{\text{beat}} = 1/f_{\text{beat}}$ which is simply the amount of time between consecutive constructive interferences.

The other feature of this graph is that as well as the variation in the amplitude, we still have a quicker oscillation. The frequency of this oscillation is called the *carrier frequency*, and it is the average of the two frequencies:

$$f_{\text{carrier}} = \frac{f_1 + f_2}{2}$$

If we are listening to sound, the carrier frequency determines the *pitch* (i.e. note) that is heard. If we are dealing with electromagnetic waves, the carrier frequency determines the colour that is seen (or not – depending on whether or not f_{carrier} lies in the visible spectrum!)

,

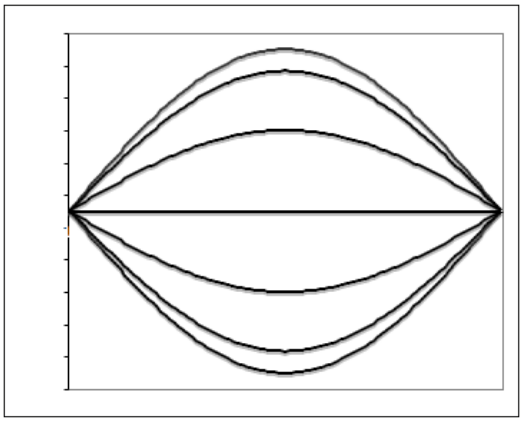
8-2-5 Standing waves

Thus far we have restricted our discussion of waves to waves that travel. In all our examples until this part, one could follow the location of a maximum and observe it moving (to the left, or the right, or outward, for instance). Another important class of waves exist called standing waves. For a standing wave, the position of the maximum and minima do not travel, but remain in place. You may have noticed standing waves when you wiggled one end of a string, slinky, rope, etc while the other end was held fixed.

We will begin our discussion of standing waves by noting what occurs experimentally when standing waves form. Initially, it may be unclear why standing waves fit into our wave unit. After all, in §8-1-1 we emphatically emphasized that waves required propagation of a disturbance. Once we establish the idea of standing waves, we will use our model of interference to make sense of them.

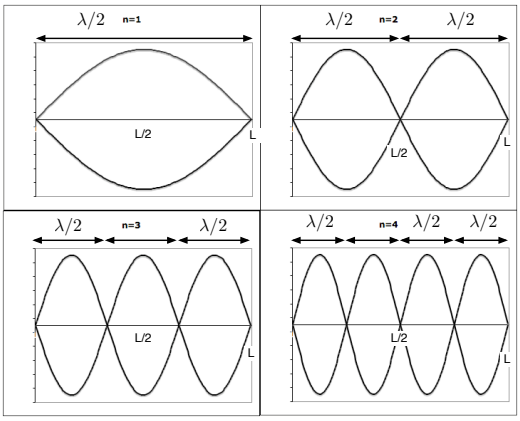
8-2-5-1 What is a standing wave?

In the most general sense, we have already defined a standing wave as a wave that does not travel. How do these waves come to exist? Imagine you have a string attached at both ends that is under tension, like a guitar string. If you try to vibrate it at a particular rate, you may or may not be successful depending on the frequency you choose. At most frequencies, the wave you start will travel to the one end intact, but upon reaching it the shape of the wave distorts and overall the string no longer appears to carry a wave. Nowhere will the string displace very far from equilibrium. Only at certain frequencies will you see a sizeable displacement. If you begin vibrating at an extremely low frequency and gradually increase the frequency, the first place the string response will result in a wave like



Each of the seven lines in the image is like a photograph of the string at a particular instant. As you know from §8-1-4 this is equivalent to a displacement versus position graph of the string, at seven different times. Notice that the displacement at both ends is zero. This makes a great deal of sense, because both ends are attached and thus cannot move. We will call any part of a standing wave that experiences no displacement over time a *node*. Also notice there is one spot in the middle that experiences maximal displacement at each time. Any spot that exhibits this behaviour will be called an *antinode*.

If we increase the frequency of our vibrations, we will lose the wave shape for awhile. The next three frequencies resulting standing waves are shown below, along with the frequency we already discussed.



In each image, the arrows highlight a distance of a half wavelength. If we use L to denote the length of the string, then for the first frequency,

$$\lambda = 2L$$

because only half of a wavelength fits on the string. For the second lowest frequency,

$$\lambda = L,$$

since an entire wavelength fits on the string. Similar relationships can be established for $n = 3$ and $n = 4$. In general, for waves on a string that are attached at both ends,

$$\lambda = 2L/n.$$

Here, the various n specify which *harmonic* we are discussing. The lowest harmonic, with $n = 1$, is called the *fundamental*.

We have now developed a relationship between harmonic (n) and wavelength (λ). If we knew the wave speed on the string, we could determine the frequency we have produced using

$$v_{\text{wave}} = \lambda f.$$

Because frequency does not change between media, whatever frequency is produced on the string is reproduced in the air and eventually makes it into our ear. It is the frequency that we hear as a particular note. To make sure the instrument plays the correct note (i.e. frequency) a musician must first

tune the instrument. To do this the musician can change the tension that the string is under (which adjusts the wave speed as discussed in §8-1-2-5). While playing the guitar a guitar player chooses different notes by putting fingers on the fretboard, which changes the available length of string to fit the wavelengths on. In general, musical instruments produce many harmonics when playing any particular note. When we name the note the instrument plays as a single frequency, such as 440 Hz, we refer to the *fundamental frequency*. The different combinations of harmonics give instruments different sounds, allowing people to differentiate between pianos, guitars, and violins even though all three instruments are stringed.

We have now explored standing waves with both ends attached in some detail. A similar analysis could be done for standing waves with only one end attached and the other end free. In this case, the attached end will still behave like a node, but now the free end will behave like an *antinode*. For instance, only one quarter ($\lambda/4$) of a wave will fit along the length of the string for the fundamental frequency.

Test yourself:

Draw the first four harmonics for a wave with one end attached and one end free. Can you determine a general relationship between the string length and the wavelength?

Applying the interference model

Now that we have some sense of what standing waves are, it is time to make sense of them. There are two independent ways of making sense of this phenomenon in terms of the interference model. We will explore both briefly. Both methods involve waves travelling down the medium in opposite directions and interfering along the way.

Reflection at the boundary

First, we will that we send a continuous wave down our medium, it hits the boundary at the end, and reflects. We now have two waves in the medium: 1) the wave we are originating, and 2) the reflected wave. The two waves will have the same frequency, as that is determined by the source. They will also have the same wavelength, since both waves travel in the same medium. Because the waves travel in opposite directions, they have different signs in

front of the position time. Mathematically, we have

$$y_1(x,t) = A\sin\left(\frac{2\pi t}{T} - \frac{2\pi x}{\lambda} + \phi_1\right)$$
$$y_2(x,t) = A\sin\left(\frac{2\pi t}{T} + \frac{2\pi x}{\lambda} + \phi_2\right)$$

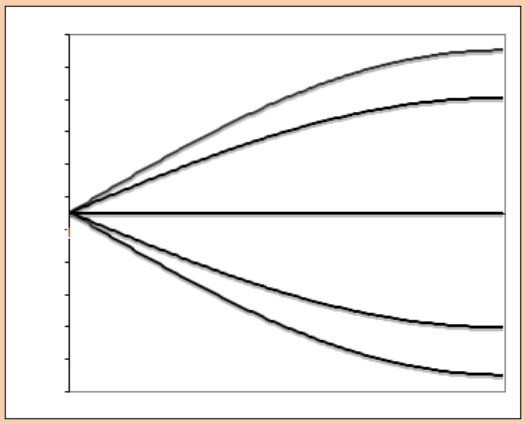
Notice that we left subscripts on the phase constant term, ϕ. When the wave hits the end of the medium, two different things can happen to the phase constant. In one case, called a *soft reflection*, the phase constant remains unchanged and $\phi_2 = \phi_1$. In the other case, called a *hard reflection*, the phase constant of the reflected is completely out of phase with the phase constant of the incoming wave, so $\phi_2 = \phi_1 + \pi$.

Example #1:

Determine if reflection at the free end of a rope is a hard reflection or a soft reflection.

Solution:

To visualise the phenomenon better, let's first sketch the situation. We could choose any harmonic, and any behaviour for the other end we want.



The sketch chosen shows the fundamental with one end attached and one end free, at five different times. At the free end of a rope, there is an antinode. At some specific times, the free end has a maximal displacement. In general, at any given instant in time, the free end has more displacement than any other part of the rope. We will keep this in mind.

As with any interference problem, there are three terms to consider that might cause interference.

1. The time term. In this case, the waves in question have the same period. There will be no total phase difference contribution from the time term.

2. The spatial term. We are examining the wave just as it turns around. Neither wave has travelled further than the other wave. There will be no total phase difference contribution from the spatial term.

3. The phase constant term. We seek to determine what the phase constant might be. We know the choices for $\Delta\phi$ are 0 (soft reflection) or π (hard reflection).

At this point, we know that the interference at the free end is entirely from the difference in phase constant, so $\Delta\Phi = \Delta\phi$. We either have $\Delta\Phi = 0$ or $\Delta\Phi = \pi$. In the first case, we would have constructive interference, and in the second case destructive interference. Clearly we are *not* seeing destructive interference, but we *are* seeing constructive. Thus, the reflection at a free end must be a soft reflection with $\Delta\phi = 0$.

A few points from the example above are worth reiterating and expanding. The free end of a rope, an antinode, is a location of a constructive interference. This location has constructive interference for all times. Though the rope itself alternates between a maximum, flat, and a minimum, the interference is always constructive. This might seem counterintuitive at first (constructive interference and zero displacement?!), but please turn back to the sketches of interference from speakers in §8-2-3-1. There are several places where the displacement of both waves is zero (total phase $\Phi = 0$ or 2π etc.), so the total displacement of the *sum* is also zero. Nonetheless, this is a spot of constructive interference.

Infinite interference waves

There is a second way to apply the interference model to understand standing waves. In this case, we do not imagine any reflections, so do not need to worry about whether a boundary change is 'hard' or 'soft.' Unfortunately, it is a bit more abstract.

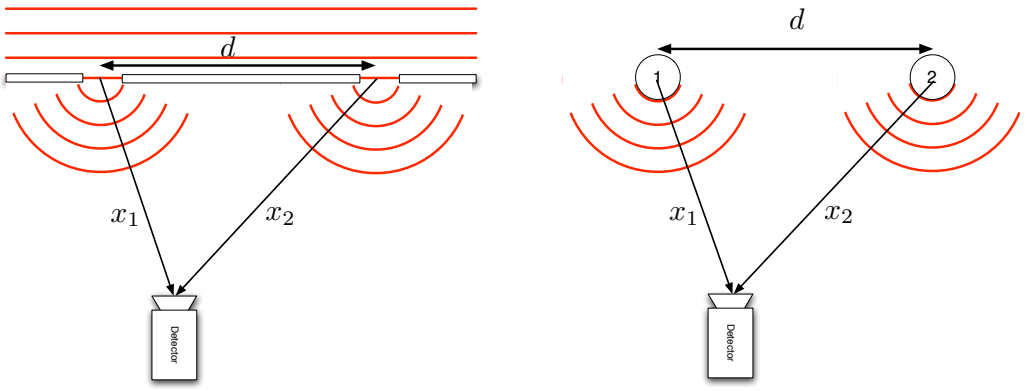
Imagine there are infinite waves travelling in opposite directions. Although we will be thinking about a small section of medium, like a length of string, we imagine that the waves extend beyond the medium in question. Throughout all of space, these waves are interfering. Their interference is like a giant

standing wave in all space that has nodes and antinodes in every direction, with no ends. To use this idea, we think about the particular type of interference we have (like both ends fixed, or node-node interference). We then take only the portion of the total interference pattern we need, apply this to our specific phenomenon, and ignore the rest.

8-2-6 Two-slit interference

When a wave passes through a wide slit, the pieces of the wave that hit the boundary stop, but the rest of the wave goes through unaffected. This effect is responsible for the shadows that we see – objects abruptly stop light, and leave a dark region. When the slit is smaller than or roughly the same size as the wavelength, something interesting happens. Instead of waves passing straight through the gap in the boundary, the wave travels outward. The key point is that (at least for slits that are much smaller than the wavelength of the wave) the slit acts like a source of waves in the region past the barrier.

If we have two slits in a screen we can get interference from the waves coming from one slit interfering with the waves coming from another slit. To figure out the sort of interference we get, we simply need to use the phase charts that we have already introduced. The most common situation we will discuss is one in which a *single* waves hits two slits at different points.



We can treat this situation like two sources with the same separation, but this special setup gives us two extra advantages:

1. Because both of these slits have the same wave reaching them, we know that the frequencies are the same.

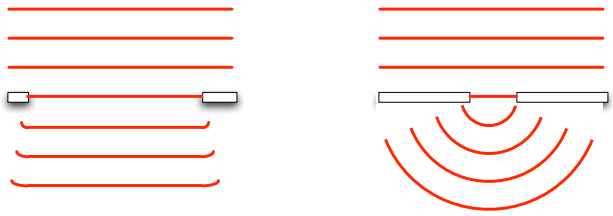


Figure 8-2.3: An example of light passing through a wide slit on the left, and a narrow slit on the right. Here wide or narrow are comparisons to the wavelength of light. The lines represent the peaks of the waves.

2. In *this special case* we have the peaks arriving at the slits at the same time. Hence we can treat the two "sources" as having the same constant phase ϕ. (This would not necessarily be true if we placed the screen with the two slits at a different angle).

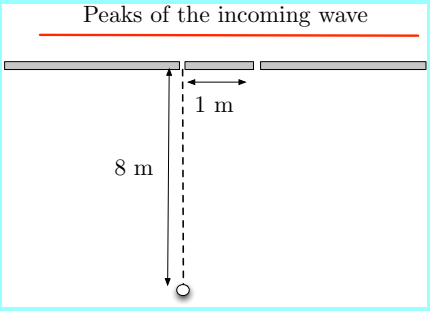
For this special case, our phase chart is

	$2\pi\frac{t}{T}$	$\pm 2\pi\frac{x}{\lambda}$	ϕ	Φ
Wave 1	$2\pi\frac{t}{T}$	$-2\pi\frac{x_1}{\lambda}$	ϕ	
Wave 2	$2\pi\frac{t}{T}$	$-2\pi\frac{x_2}{\lambda}$	ϕ	
Change	0	$-2\pi\frac{(x_1-x_2)}{\lambda}$	0	$\Delta\Phi$

$$\Rightarrow \Delta\Phi = -2\pi(x_1 - x_2)/\lambda$$

Example #2:

A speaker in the distance is playing a single note. The peaks of the waves hit the wall together, and pass through the two holes separated by 1 m. At different locations behind the wall it is either quiet or soft. What can we tell about the frequency of the note being played if we know that the location of the hollow circle is quiet?



Hint: you will need to know that the speed of sound is roughly 340 m/s, and that the range of human hearing is 20 – 20,000 Hz.

Solution:

We know that the two waves will have the same frequency, and as the peaks are arriving together they must also have the same constant phase ϕ. Plugging this into the phase chart like before we find

$$\Delta\Phi = -2\pi\frac{\Delta x}{\lambda}.$$

Because we know this is a quiet location we know that

$$\Delta\Phi = -(\text{odd})\pi.$$

From the geometry of the problem we can calculate Δx. The distance from the slit on the left to the hollow circle is 8 m. The distance to the right slit can be found by using Pythagoras's theorem:

$$x_{\text{right}} = \sqrt{(8\,\text{m})^2 + (1\,\text{m})^2} = \sqrt{65}\,\text{m} \approx 8.062\,\text{m}.$$

Therefore Δx is given by

$$\Delta x = (8.062 - 8)\,\text{m} = 0.062\,\text{m}$$

Putting the fact that $\Delta\Phi = -(\text{odd})\pi$ and $\Delta x = 0.062\,\text{m}$ we have

$$(\text{odd})\pi = 2\pi\frac{0.062\,\text{m}}{\lambda} \Rightarrow \lambda = \frac{2\Delta x}{\text{odd}} = \frac{0.124\,\text{m}}{\text{odd}}.$$

This gives us information on the wavelength, to get to the *frequency* we use the fact that the speed of sound is $v_{\text{sound}} = 340\ \text{m/s} = f\lambda$. The frequencies the speaker could be playing are

$$f = \frac{v_{\text{sound}}}{\lambda} = 2740 \times (\text{odd})\,\text{Hz}$$

We cannot pick a unique answer, only find *possible* answers. We can list all the possible answers if we know that humans cannot hear beyond 20,000 Hz (and thus the speaker would not be loud); otherwise an acceptable solution would be to stop here. Plugging in values we have (to three sig. fig.)

Odd number	**Frequency**
1	2740 Hz
3	8230 Hz
5	13700 Hz

Odd number	**Frequency**
7	19200 Hz
9	24700 Hz

i.e. only possibilities that people can hear are 2740 Hz, 8230 Hz, 13700 Hz and 19200 Hz.

8-2-6-1 Approximating the path length difference Δx

As the last example showed finding the path length difference exactly can be quite lengthy. Fortunately if we are interested in the type of interference a long way from the slits we can use an approximation that is considerably easier. We picture the setup shown in figure 8-2.4. Here we are interested at the type of interference we get at the location of the white circle located a distance y from the center on a screen a distance L from the slits. The approximation that we are going to use will involve the angle θ the angle between the dotted lines. If we wish to relate this back to y and L we may use trig: $\tan\theta = y/L$.

We start by drawing part of a circle with its origin on the hollow dot and that goes through the closest slit as shown below on the left.

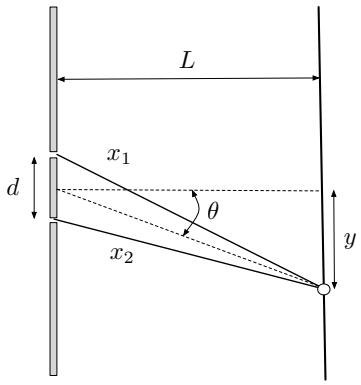
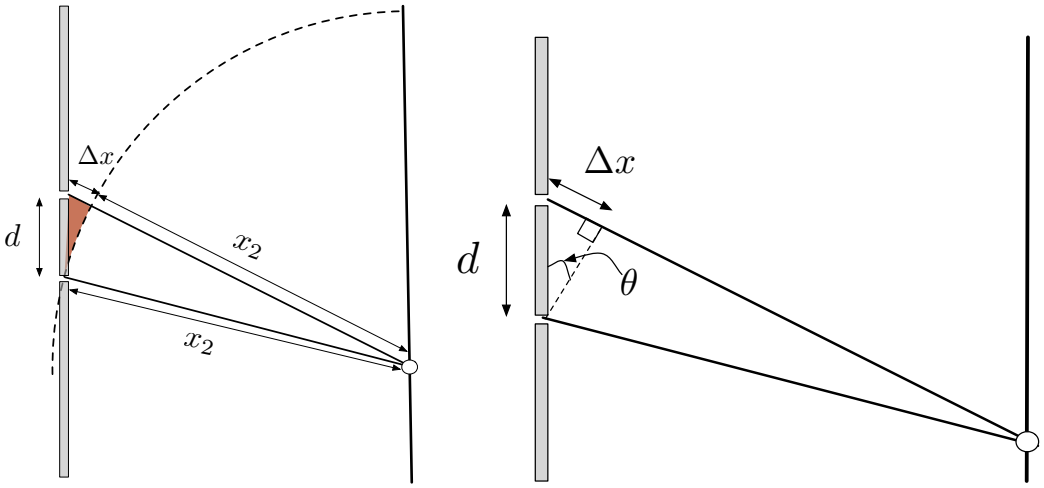


Figure 8-2.4: Defining the variables for a common two-slit interference problem.



Because this is a circle we know that *all* points on the circle are an equal distance (namely x_2) from the hollow point. The path length difference Δx is precisely that small distance that lies between the edge of the dotted arc and the furthest slit. Here comes the approximation: if the circle was instead a straight line, the shaded region would be a right-angled triangle. Because we are dealing with an arc of a circle it is *slightly* curved this is not exact. As long as the hollow circle is a long way away ignoring the curvature of the circle is not a bad approximation. Once we make that approximation we are lead to the simpler picture above on the right. Looking at this right-angled triangle we have

$$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{\Delta x}{d}.$$

Rearranging we find that $\Delta x = d \sin\theta$. We must remember this relationship is only approximate as the triangle is not precisely a right-angled triangle.

Test yourself:

Prove that the angle θ shown in the previous diagram on the right is the same angle θ in figure 8-2.4.

8-2-7 Summary

This section did not involve much in the way of new material; most of it was exploring the implications of the idea of superposition:

- Superposition is the way of combining the effects of two (or more) waves.
- To superpose two (or more) small waves, we add together the effects of the wave on a particular point at a particular time.

For mechanical waves "effects" means the displacement of the medium Δy. For sound waves "effects" can refer to either the change in pressure ΔP or the displacement of the atoms in the medium Δy. For light waves "effects" refers to the magnitude of the electric or magnetic field. Writing the second bullet point for two sources mathematically

$$\begin{aligned}\Delta y_{\text{tot}}(\text{at location } x, \text{ time } t) = {} & \Delta y_1(\text{at same location } x, \text{ same time } t) \\ & + \Delta y_2(\text{at same location } x, \text{ same time } t).\end{aligned}$$

Interference is a "taxonomy of types of superposition". Specifically we introduced

- Constructive interference: where the two waves added together maximally. For harmonic waves this implies that they are in phase, or out of phase by $\pm 2\pi$, $\pm 4\pi$, $\pm 6\pi$, ...
- Destructive interference: where the waves cancelled each other maximally. For harmonic waves this implies that they are in phase, or out of phase by $\pm\pi$, $\pm 3\pi$, $\pm 7\pi$, ...
- Partial interference: anything that is not constructive or destructive interference.

- Three terms contribute to the interference condition. We need to consider the combined effect of a possible path length difference (Δx), the effect of the sources being in phase or not ($\Delta\phi$) or that the frequencies are different. The actual type of interference only depends on $\Delta\Phi$.
- The phase chart, which is a way of organising the three terms mentioned above.

8-2-8 Exercises

Spot the error

Each of these numbered explanations contains at least one imprecise or incorrect statement. Can you find the error and help the misguided student?

1. "When two waves interfere, the wave only has zero displacement from equilibrium at a location of destructive interference. At a location of constructive interference, the displacement from equilibrium is always maximum."
2. "Constructive interference occurs whenever two sources of the same frequency are separated by an integral number of wavelengths"
3. "Constructive interference occurs whenever the path length difference is an integral number of wavelengths for waves of the same frequency."
4. "If $\Delta\Phi = 0$ we have neither constructive nor destructive interference because 0 is neither even or odd." (You should try to correct both the mathematical statement *and* be able to tell on common sense arguments whether the interference is constructive, destructive or partial.)

Unit 8:

8-3: Geometric optics

In Physics 7C so far we have dealt with the introducing waves and their interactions. Our model of waves has been so useful because we have been able to use the same basic ideas to a wide variety of phenomena; namely waves on ropes, sound waves, light waves and other types of waves. In this chapter we will be dealing with waves that travel from one medium to another. In such a case two things can happen: part of the wave can bounce back into the original medium which we refer to as *reflection*, and part of the wave can get into the next medium (transmission). When a 2-D or 3-D wave travels into a new medium the wave is typically "bent", a phenomenon we call *refraction*. We can combine these effects of reflecting and bending waves to make the waves appear as if they are being created at different locations than they actually are. If the waves in question are light waves then this means that we see "images" at places distinct from where the objects themselves are. While most of our examples will involve light it is important to realise that all types of waves will reflect and refract as they pass from one medium to another. (For completeness we mention that there are two other methods by which the path of light can be altered: absorption and scattering. We will not develop these further.)

So far when picturing waves we have thought of the oscillating sine function. For 1-D waves this was an adequate way of picturing the wave. For higher dimensional waves this representation becomes difficult, and so we introduce the idea of *wavefronts* and *rays*.

8-3-1 Rays and wavefronts

Let us start by thinking of dropping a stone in water and letting the ripples propagate outward. Some time later we may photograph the wave we have created as shown on the left of the figure 8-3.1. In this figure, parts of the wave are obscured and it is generally difficult to draw and visualise

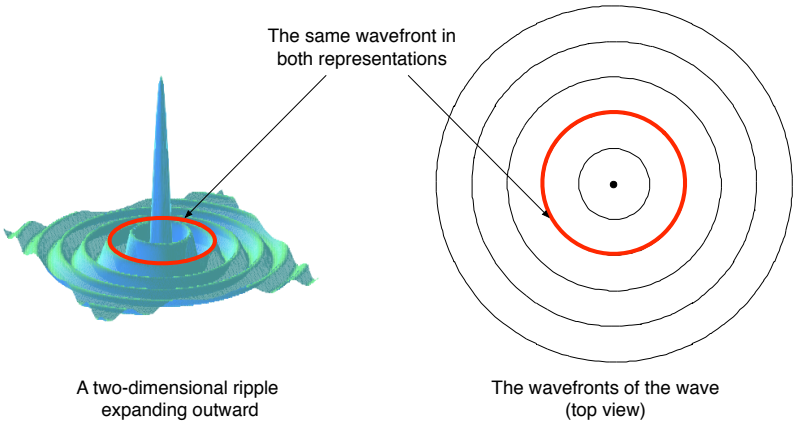


Figure 8-3.1: An illustration of the surface of water it has been disturbed by dropping a stone on the left. The picture on the right is a top down view showing the wavefronts leaving. Here we have chosen the wavefronts to be the peaks of the wave. The red circle indicates the same peak in both pictures.

.

interactions between waves. These limitations make it a difficult to use representation, so we adopt a more useful representation that takes some of the details out. One such representation is the *wavefront* representation in which we choose to only draw one part of the waves. Because we know the wave is oscillating up and down, by looking at the picture of the wavefronts you should have a reasonable idea of what the wave is doing. Occasionally we shall draw wavefronts for the peaks and troughs in different colours so that we can superpose them, recalling that peak + peak or trough + trough will give *constructive interference* while peak + trough will give *destructive interference*.

If the medium that the wave is propagating through is isotropic (i.e. the same in all directions) the wave will spread out at the same speed in all directions and the wavefronts will be concentric circles (for 2D waves) or concentric spheres (for 3D waves). As we get a long distance from the source, we only see a small portion of the wavefront (shown in bold in figure 8-3.2) and this part of the wavefront looks almost flat[1] When the waves are (approximately) flat we call them *plane waves*, because the wavefront then

[1]This is for the same reason the Earth appears flat. Even though we know the Earth is spherical we only see a little part of it. A little part of a circle looks almost flat.

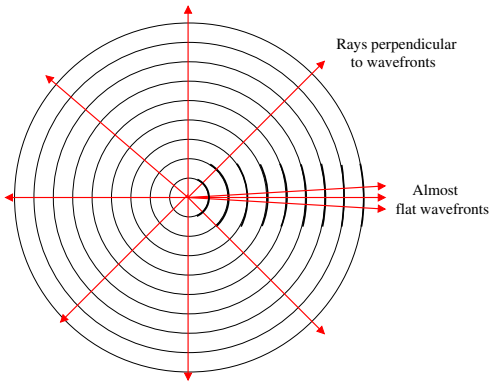


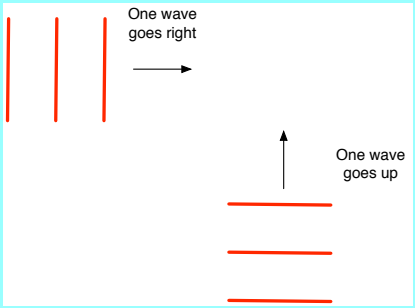
Figure 8-3.2: A picture showing how the wavefronts look like flat planes at a large distance from the source. Notice that the rays from the bolded region are also approximately parallel.

resembles a plane (at least if the wave is three-dimensional).

In addition to thinking about the wave, we can think about the direction that a particular piece of the wave is travelling. We can join these directions and they trace out a path of a particular piece of a wave. These paths are called *rays*, and are always perpendicular to the wavefronts. Examples of rays are shown in the figure as arrows. Notice that we can draw whichever rays are convenient to use. In the pond example we drew many rays going in all directions, but concentrated the rays in the part where we discussed the wavefronts looking flat. This is because we would also like to demonstrate that the rays are close to parallel a long way from the ripple. We are *not* saying that there is more energy in that part of the wave than any other, we are simply drawing those rays because they bring attention to the phenomenon we wish to discuss. Throughout this section on optics we will select our rays to illustrate the points we wish to make.

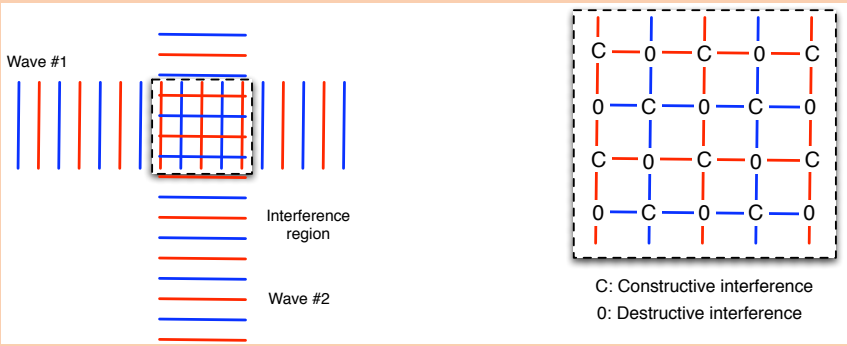
Example #1:

Two (plane) waves with the same wavelength are coming in from far away. One is travelling right and the other is travelling from the bottom of the page to the top, as shown in the picture below. Find the interference pattern created by the two waves when they cross.

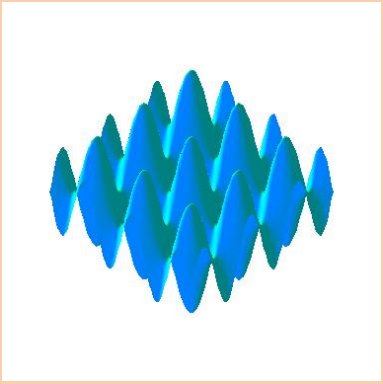


Solution:

The picture above only showed the wavefronts for the peaks. We are going to introduce the wavefronts for both the peaks (in red) *and* the troughs (in blue). We know if two of the same colour cross (i.e. peak + peak, or trough + trough) we get constructive interference, while if a red and blue cross we get destructive interference. Drawing the picture we have



On the left we have drawn the individual wavefronts crossing, and on the right we have put C for constructive interference and 0 for destructive interference. Both of these are easier to visualise than the picture of the wave keeping all the information shown below:



8-3-2 Optics and images

Now that we have the concepts of *rays* and *wavefronts* we move on to the subject of (geometric) optics. The approximation that geometric optics makes is that the rays travel in straight lines until they hit a surface. When the ray encounters a surface it can either bounce back (reflect) or bend (refract) but then travels on in a straight line.

When is geometric optics applicable? If light started heading toward a two-slit experiment we have learned in the previous section that the light would diffract i.e. we would get bright and dark fringes. But this idea that light travels in straight lines would tell us we would just get two bright bands! It is important to realise that the idea of the rays travelling in straight lines is *only* valid if the wavelength of the light is much smaller than any of the objects or slits the light will encounter. In terms of the diffraction problem, we are saying if the slit is much greater than the wavelength of light then we can ignore diffraction. Mathematically the approximation we are making is

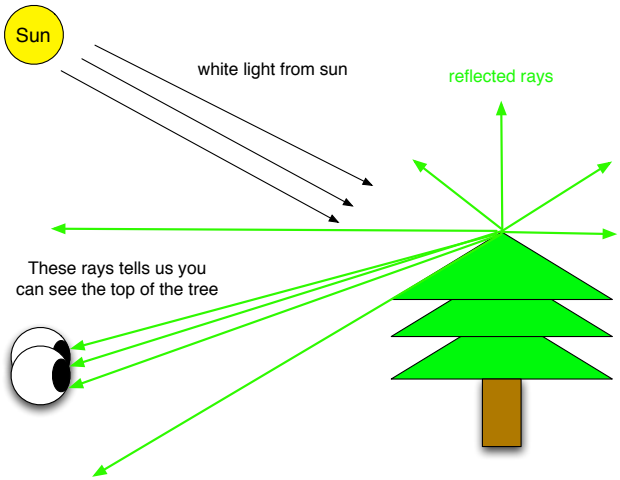
$$\lambda \ll d, \qquad \text{(geometric optics approximation)}$$

where d is the size of any slit or object the light encounters, and λ is the wavelength of the light.

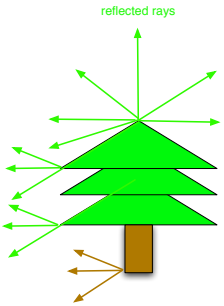
Before going too much further it is worth considering how we see things. We are not going to deal with the eye directly yet, as the eye is a complicated system. We will deal with some aspects of how corrective lenses work a little later in section §8-3-3-4 but for now all we need to know about the eye is that it can only see the light rays which reach it. Let us think for a moment about how we actually see an object such as a tree. The tree does not give off visible light – we can tell this by the fact that on a dark night we cannot see a tree. On a bright day we see the tree because the sun gives off light, which hits the tree and reflects in a lot of different directions. The light that is reflected is *not* the same as the light that comes in – otherwise everything outside would be the colour of the sun! Instead objects reflect certain colours preferentially, and absorb others. The tree leaves, for example, reflect green strongly but absorb most of the other colours. When we shine sunlight (which is a combination of all the colours) the green is strongly reflected which makes the leaves appear green to us[2]. We can "see" a particular point on the tree

[2]The green leaves also absorb almost all red light. If you were to go into a photographer's room with only red bulbs with a green leaf it would look black, as there is no green light in the room to reflect back at you!

if some of the rays that reflected from that point enter our eyes. The act of seeing only the top point of the tree is summarised in the picture below:



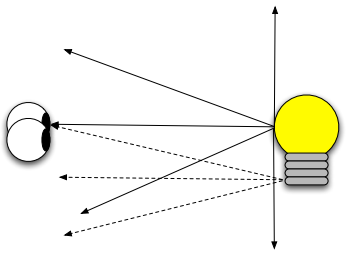
Here multiple rays have been drawn that enter the eye. Our diagram is not meant to suggest that a disproportionate amount of light enters the eye; we draw a higher density of light rays in this region because we are more interested in light that reaches the eyes than what happens to light going in other directions. Because multiple rays enter our eyes at slightly different angles our brain can judge how far away the top of the tree is from us. In doing this, our brain assumes that the light rays travelled to us in a straight line. This is not only being done for the top of the tree, but for *every* point on the tree!



The phenomenon of light scattering in all directions when it hits an object is called *diffuse reflection*. We will come back to how this comes about when we discuss reflection in §8-3-2-1.

The story for a luminous object (i.e. one that emits light rather than reflects it) is not that different. Consider a lightbulb. It has rays going off in all different directions, and we can see the lightbulb if some of those rays

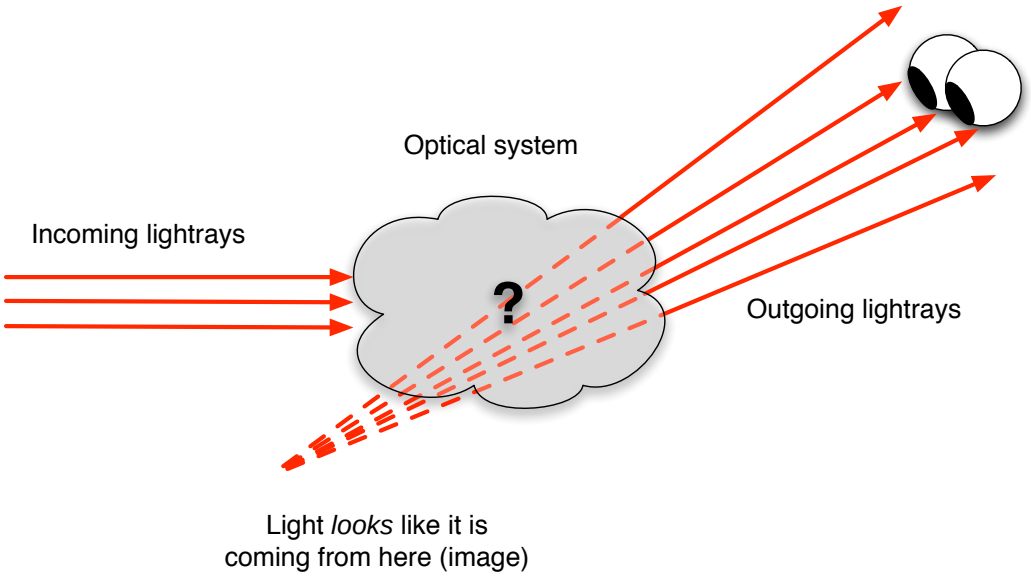
enter our eyes. Our brain tells us where the lightbulb is by assuming that the light rays travel in straight lines.



Light rays from a luminous point on the lightbulb (solid)
Light rays from a non-luminous point on the lightbulb (dashed)

For the purposes of figuring out where something is or how the light rays travel, it is not important whether the object creates its own light, or if the light is merely reflected; in both cases the object has light bouncing off it at all angles.

These considerations about how we see things raise an interesting possibility. The only information we have access to for sight is the light that reaches our eyes. If we can bend or twist the path light takes, then we will judge objects to be at different places. This is exactly what happens when we look in a mirror and see an image of ourselves! Our study of optics is essentially the study of how light given off by objects (whether this light is created by the object or simply reflected) can be manipulated into appearing like it comes from somewhere else. We call this somewhere else an *image*.

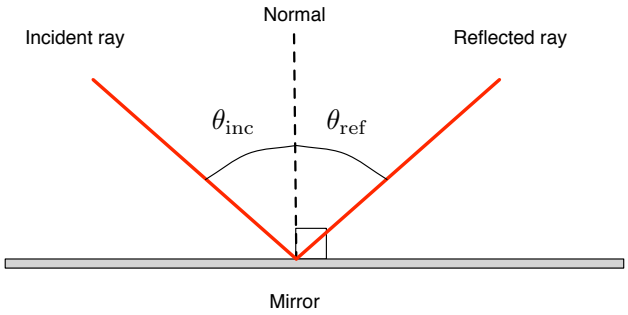


The dotted rays are the ones our brain traces back, while the solid rays are the actual light rays. We see that to locate where the image of a point is we must look at *multiple rays* from the *same* point on the object, and see where those rays appear to come from. To figure out the entire image of an object then we must find the image of each point on the object individually. Before constructing the idea of images too much further, we shall consider the two simplest ways in which we can change the path of a light ray: reflection and refraction.

8-3-2-1 Reflection

When a wave reaches the interface between two different media, typically some of the wave will bounce back. This process is known as *reflection*. The example that most people think of when hearing about reflections are optical reflections in mirrors. Another familiar example of reflection comes from water waves. As the water waves travel they will reflect off objects that are floating on the water, and also reflect off the walls of the container holding the water. Most of us are familiar with the concept of *echoes*, which are the reflections of sound waves. In fact, any kind of wave can undergo reflection!

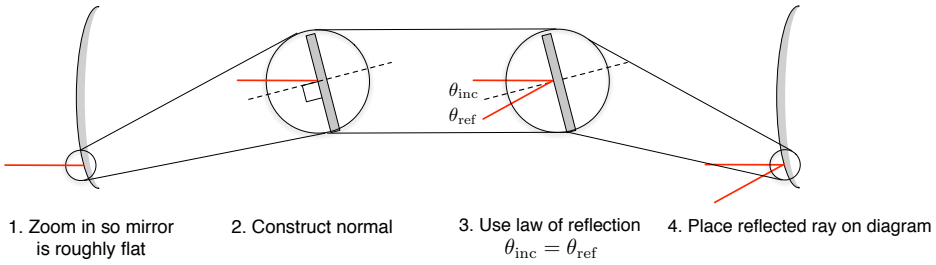
Let us start by describing how light bounces off a mirror. If you have ever shone a torch or laser pointer at a mirror in a dark room, you may have noticed that the light reflects in a particular direction instead of scattering everywhere. If not, you will get some experience with this in one of your physics labs. One of the questions we would like to answer is "which direction does the light bounce off the mirror?" The answer to this question depends on the angle the light hit the mirror, and in order at answer it precisely we need to introduce the concept of a *normal*. For a flat mirror, the normal to the mirror is just the line that pokes "directly out of the plane of the mirror". This means that the normal will always be at 90° to the mirror. If we shine a ray of light onto a mirror, the angle between the ray and the normal is denoted θ_{inc}, where "inc" is short for *incident* ray. The ray of light that bounces off the mirror is on the other side of the normal, but at the same angle! The picture below may make this more clear:



The angle that the light comes off at is sometimes called θ_{ref} where "ref" stands for *reflected ray*. The *law of reflection* tells us that[3]

$$\theta_{\text{inc}} = \theta_{\text{ref}}$$

To deal with bent mirrors, such as fun-house mirrors, we apply almost exactly the same principle. The only difference is we look at where the light hits the mirror, and at least in a small region that part of the mirror is almost flat. We can then find the normal and use the law of reflection to find where the light ray goes. By repeating this for many light rays we can find out anything we want to know about a curved mirror. A quick sketch of how to do this is presented below.

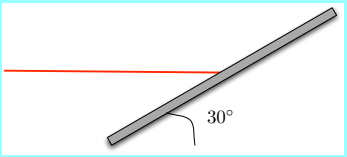


Instead of drawing the magnified image, we can also consider drawing a *tangent* to the surface at a particular point. As you should know from your mathematics class, the *normal* is perpendicular to the tangent line. You should spend some time considering the definition of a tangent line to figure out why this "magnification procedure" and "tangent procedure" are both valid ways of locating the normal.

[3]You may think that it should be $\theta_{\text{inc}} = -\theta_{\text{ref}}$, as the rays are on opposite sides of the normal. This would make perfect sense, but because the directions are usually intuitive the convention in optics is to treat all the angles as positive.

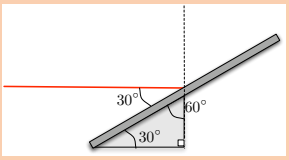
Example #2:

We have a light ray that hits a mirror as shown. The mirror is slightly tilted. Where does the reflected ray go?



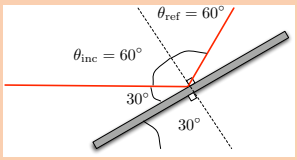
Solution:

In this problem we have *not* been directly given the angle of incidence directly, rather we will need to do some geometry to figure out what θ_{inc} is. One way of doing this is to construct a right-angled triangle using a vertical line as shown:



Because we know the sum of all the interior angles in a triangle sum to 180° we know that the angle in the upper right hand corner of the triangle must be 60°. Because the original light ray is coming in horizontally and the dotted line is vertical, we can deduce that the angle between the mirror and the light ray is 30°.

Now we introduce the normal as a dotted line. Because the angle between the mirror and the normal must be 90° we can deduce that $\theta_{\text{inc}} = 60°$. By the law of reflection we must also have $\theta_{\text{ref}} = 60°$. The final part of the solution is sketched below.

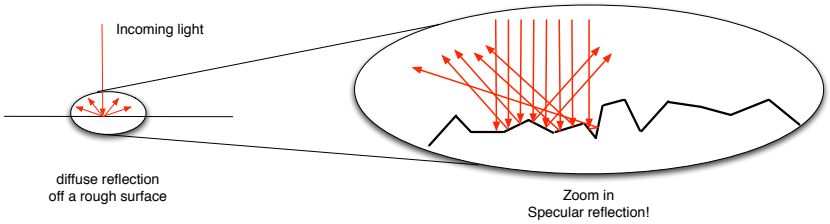


The difficult part of this problem is not the law of reflection itself, but rather the geometry of finding the incident ray. You may notice that we could have equally well used the fact that the angles between the mirror and the ray are the same for the incident and reflected ray, saving us some geometry steps. We choose not to emphasise this because when we introduce the more

quantitatively complicated *refraction*, it is important to make sure you are using the angle between the normal and the incident (or reflected) ray.

Diffuse reflection

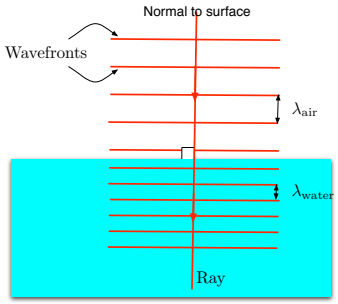
When we discussed forming images, we made a point of stating that light typically reflects in all directions, and labelled this phenomenon *diffuse reflection*. This behaviour seems very different from the idea of a light ray coming in and bouncing off in a specific direction as described by the law of reflection. The difference is that most materials are rough, and so different places have different normals. A beam of light is really a collection of many different rays, and even though the incoming rays are parallel they hit different places, and are hence at different angles of incidence from the normal. As a consequence, the outgoing reflected rays are not parallel.



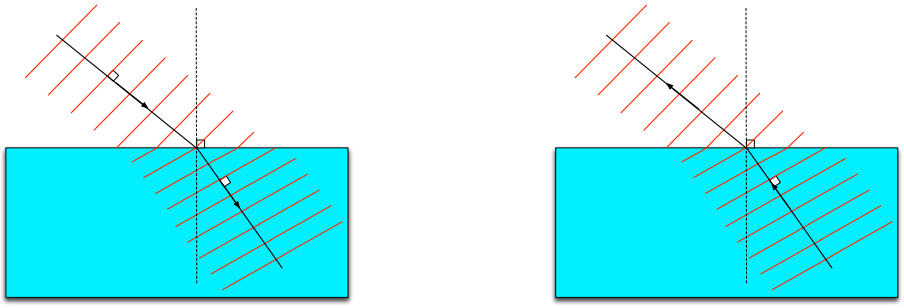
8-3-2-2 Refraction

We have already discussed that the speed of a wave depends on its medium, and this is true for light as well. Because the frequency of the light cannot change (recall that the frequency of a wave is set by the source) the period between peaks must stay the same. Because we know $v_{\text{wave}} = f\lambda$ we see that if v_{wave} changes λ must also change. One easy way of understanding this is if v_{wave} is small, then the wave cannot travel very far in one period so λ is small, and if v_{wave} is large then one peak can travel further in a period so λ is large.

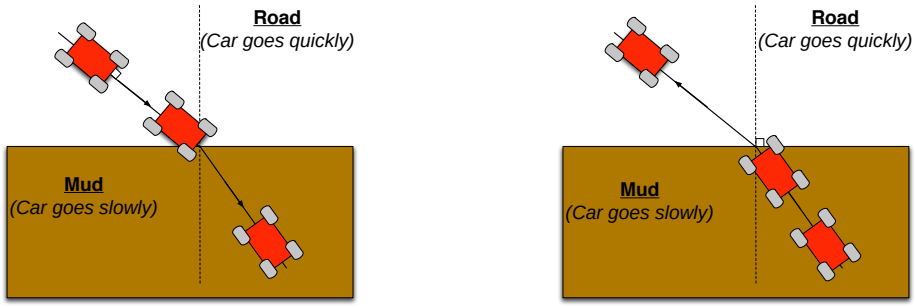
Let us consider the concrete case of light travelling from air into water, and inform you that light travels faster in air than it does in water. Consider the case of a plane wavefronts travelling from air into water, and the wavefronts are parallel to the water surface. This is called *normal incidence*, as the light rays are travelling along the normal of the air-water boundary. We note that light travels faster in air than it does in water, and this makes the wavelength of light in water shorter. Even so the path that the light takes is unaffected



The situation becomes much more interesting if the wavefronts of the light rays are *not* lined up exactly with the air-water boundary as shown below.



In this case part of a wavefront enters the water and slows down, while the rest of the wavefront stays in the air at its faster speed. The wavefront in the air "overtakes" the wavefront in the water, but they still have to join smoothly at the boundary. This causes the whole wavefront to bend. A useful analogy is the idea of your car getting stuck in mud: if one tire goes forward faster than the other, this causes your entire car to turn. Note that this idea works for both the cases where the car goes from a fast medium (e.g. the road) to a slow one (e.g. the mud) **or** travels from the slow medium to the fast one.

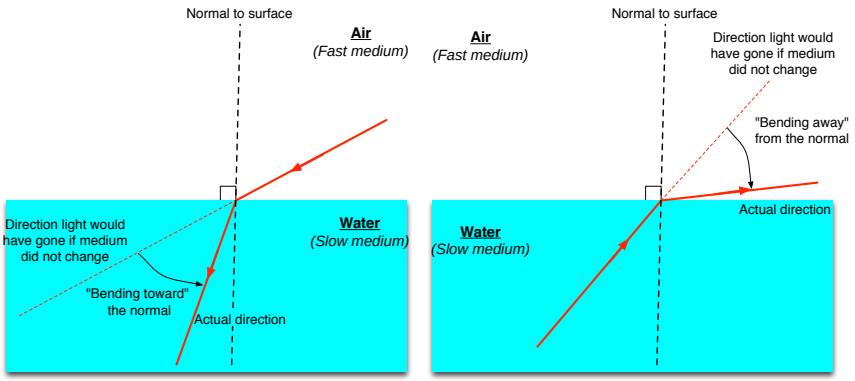


This time the rays have been indicated on the wavefront picture in black, and we can see the ray noticeably bends as we go from air to water. There is

nothing particularly special about air or water; this bending occurs for any two different media where the waves have different speeds. This bending of light as it goes from one medium into another is called *refraction*. Notice that within each medium the light rays travel in straight lines.

You may wonder when a light ray hits a surface, how can we tell if it is going to be reflected or refracted? The answer is that a light ray is typically *both* reflected and refracted. We already have some familiarity with this from our experience with swimming pools. It is possible to see the sun from inside a swimming pool, so we know that light from the sun must be able to make it into the water. Therefore the sun's rays are refracting as they enter the water. Someone standing on the side of the pool can also see the reflection of the sun on the water's surface (usually referred to as "glare"), so the sun's rays must also be reflecting off the surface of the pool. There is no contradiction here – when the sun's rays hit the surface a ray reflects and another ray refracts. This does not violate the conservation of energy, as each of the rays will have less energy than the incoming ray. *Remember that the rays do not signify a specific amount of energy.* At the moment we are simply concentrating on the refracted ray and omit the reflected ray from our discussion, but it is there nevertheless.

In refraction it is common to talk about the "fast" medium (the medium with the greater wave speed) and the "slow" medium (the medium with the lesser wave speed). In the case of air going from air to water, the fast medium is air and the slow medium is water. The examples of refraction showed that light that goes from a fast medium (e.g. air) to a slow medium (e.g. water) that the light ray bends toward the normal. As an exercise you should be able to show that as light travels from a slow medium to a fast medium the light rays bend *away* from the normal. The diagram below illustrates precisely what is meant by bending away or bending toward the normal.



That is all the qualitative information we need about refraction. We now turn to the *quantitative* task of determining precisely which way a refracted ray travels as it goes from one medium to another. We know that the amount of bending depends on the speed of the wave in the medium. For convenience we define the reactive index n for light in a particular medium as

$$n_{\text{medium}} = \frac{\text{speed of light in vacuum}}{\text{speed of light in medium}} = \frac{c}{v_{\text{wave}}}$$

The reason for this is that the speed of light in materials is typically 10^7 – 10^8 m/s, while the n for most materials is between one than five. The utility of the refractive index is that the values of n are easier to use that the values for v_{medium}. From the definition of the refractive index, we know three things:

1. $n_{\text{medium}} \geq 1$, because nothing can travel faster than the speed of light in a vacuum.

2. A fast medium has a *smaller* value of n, a slow medium has a *larger* value of n.

3. $n_{\text{vacuum}} = c/c = 1$.

The refractive indices for other materials are given in table 8-3.1.

Material	$n = c/v_{\text{medium}}$
Vacuum	1.0 (exact)
Air	1.0003
Water	1.33
Glass (crown)	1.50–1.62
Glass (flint)	1.57 – 1.75
Silicon	3.5
Germanium	4.0
Diamond	2.42
Eye	1.33
Eye lens	1.41

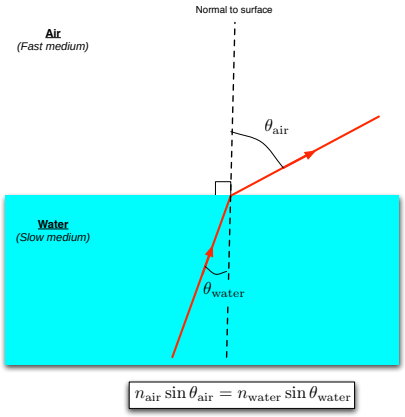
Table 8-3.1: Common refractive indices

With the definition of refractive index we can now give a *quantitative* description of refraction. We will call the refractive index in one of the media n_1 and the angle of the light ray in that medium is θ_1, and for the

second medium we will use n_2 and θ_2. All these quantities are related by Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

For a specific situation it is probably useful to ignore the labels "1" and "2" and think instead of the names of the media. This way it is much clearer which angle is in which medium. An example of this rewriting for the specific case of air and water is presented below:



Some examples of using Snell's law are given below. Typically the hardest thing about refraction problems is getting the geometry right; this is a good time to make sure that your trigonometry is under control!

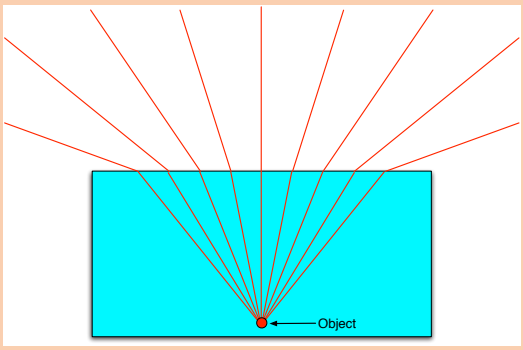
Example #3:

If we placed a point source of light in a calm pool, how would the light bend coming into the air?

Solution:

We know that rays come off the light source in all different directions. Here we have chosen to sketch a few of those directions. We know that the light ray that is at normal incidence ($\theta_{water} = 0°$) will pass straight through. We can see this from either Snell's law, or by realising that the "tyres" of the car will hit the "road" at the same time, so no bending will occur.

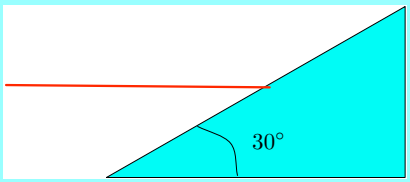
By applying Snell's law *or* by thinking about the analogy of the car wheels stuck in the mud we know that as we move away from normal incidence to higher angles, the bending becomes more severe. We illustrate this in the picture below.



You may wonder if we can ever "run out of room" as we refract – a question we will address again in §8-3-2-3

Example #4:

A ray of light (in air) comes in horizontally and hits a glass prism. The glass has a refractive index of 1.5. What angle does the light refract at *inside* the glass?



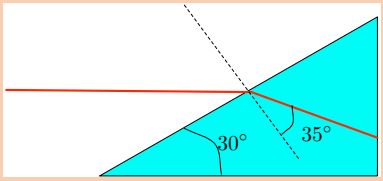
Solution:

We know from Section 8-3, ex. #2 that the incoming light ray is at an angle $\theta_{\text{air}} = 60°$ from the normal, as the angles of the tilted surfaces are the same. The refractive index for air is one, so we can use Snell's law:

$$n_{\text{air}} \sin\theta_{\text{air}} = n_{\text{glass}} \sin\theta_{\text{glass}}$$

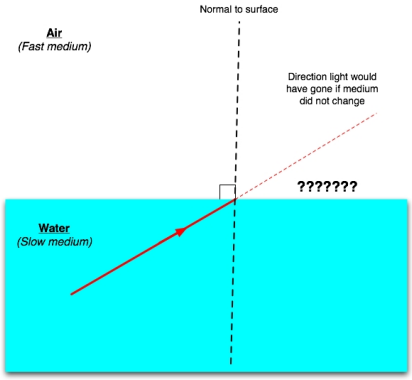
$$\Rightarrow \sin\theta_{\text{glass}} = \frac{n_{\text{air}}}{n_{\text{glass}}} \sin\theta_{\text{air}} = \frac{1}{1.5} \times \sin 60° = 0.58$$

We can use our inverse sine button on our calculator to find $\theta_{\text{glass}} = 35°$. The path of the ray looks like:



8-3-2-3 Total internal reflection

Let us do another example of Snell's law. We are going to look at light starting under water ($n_{\text{water}} = 1.33$) and being directed into the air above ($n_{\text{air}} = 1$). If we are underwater we may choose θ_{water} to be *any* angle between $0°$ and $90°$ by pointing our light source in the appropriate direction. Let us choose $\theta_{\text{water}} = 60°$ as shown in the diagram below:



We now use Snell's law to determine the angle θ_{air} that the outgoing ray will emerge.

$$\begin{aligned} n_{\text{water}} \sin\theta_{\text{water}} &= n_{\text{air}} \sin\theta_{\text{air}} \\ (1.33)\sin 60° &= (1)\sin\theta_{\text{air}} \\ \Rightarrow \sin\theta_{\text{air}} &= 1.152 \end{aligned}$$

This is obviously a problem, as we know that the sine of any (real) angle is between 1 and -1! Here Snell's law does not give us an answer.

To get a slightly more intuitive feeling for the problem, recall that if we go from a slow medium to a fast medium the light "bends away from the normal". As we increase the angle in the slow medium, we eventually "run out of room" to bend the light in the fast medium. If we shone light from air to water the refracted ray bends "toward the normal" and we would never run into this problem. A more direct way of understanding what is happening is that it takes such a long time for a piece of a wavefront to make it out of the water (where they travel slowly) into the air that any part of the wavefront that made it into the air would be forced to leave the water wavefront behind.

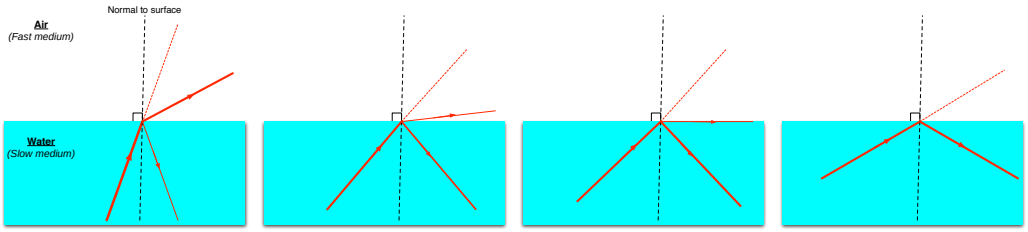
So we can understand why Snell's law does not work anymore, but if the light cannot refract what happens to it? The answer is that *all* the light

gets reflected instead. This is called *total internal reflection*, and it occurs whenever Snell's law no longer makes sense. Notice that we cannot get total internal reflection going from a fast medium to a slow one as there is always room to bend *toward* the normal. The smallest angle for which total internal reflection occurs is called the *critical angle* θ_c. To find θ_c we want to know when the refracted ray has bent as far as it can go – namely $\theta_{\text{fast}} = 90°$. For the case of water-to-air, we can find the critical angle as follows:

$$\begin{aligned} n_{\text{water}} \sin\theta_c &= n_{\text{air}} \sin 90° \\ (1.33) \sin\theta_c &= 1 \\ \Rightarrow \theta_c &= 48.8° \end{aligned}$$

i.e. for $\theta_{\text{water}} < 48.8°$ some of the light makes it into the air, but for $\theta_{\text{water}} > 48.8°$ all the light is reflected back into the water. A picture from a swimming pool demonstrating total internal reflection and the critical angle can be found in figure 8-3.3.

We do not wish to give the impression that this is a sudden change. When the light rays are perpendicular to a surface almost all the light is transmitted or "refracted". As we make the light rays hit at higher angles of incidence the amount of light refracted decreases, and the amount of light *reflected* increases. Once the angle of incidence is greater than or equal to the critical angle θ_c none of the light is refracted; instead it is *all* reflected. A sequence of images that demonstrates this change graphically, where the thickness of the line represents its brightness, is shown below:



8-3-2-4 Images again

Now that we have discussed two of the ways in which light rays can change direction, let us ask what the implications are for how we perceive things. Let us consider an object that is underwater, and ask how someone standing above would see it. In example Section 8-3, ex. #3 we already drew how the refracted rays would look. Our brain assumes that these rays are travelling

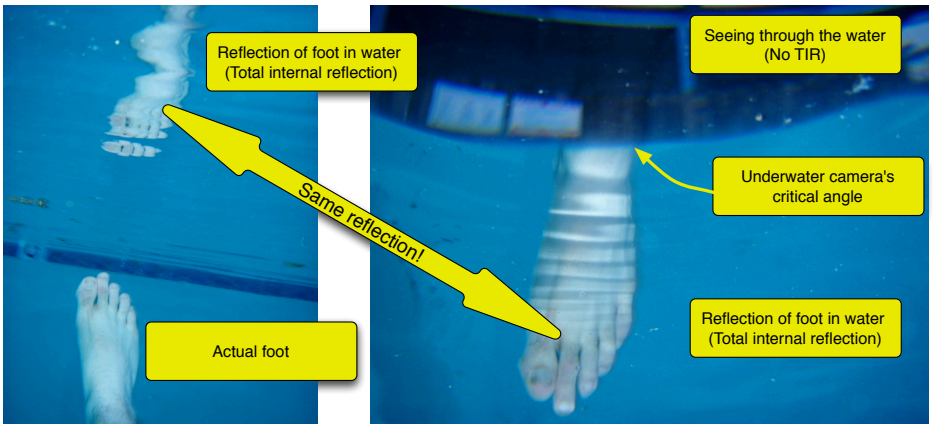
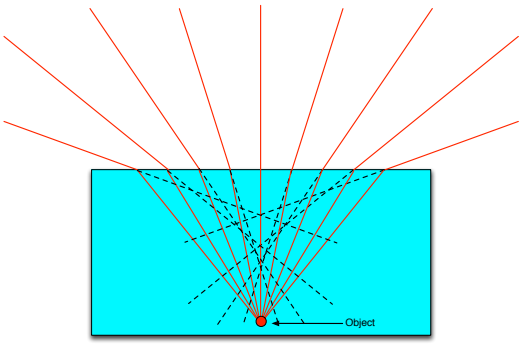


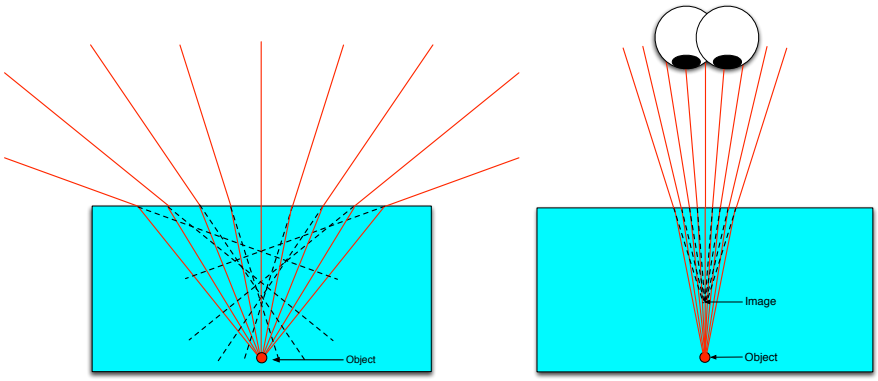
Figure 8-3.3: A picture from an underwater camera. This shows the reflection of the photographer's leg from total internal reflection, and then a distorted view of the pool house. The critical angle is the angle at which the pool house and things outside the pool can no longer be seen and only the reflection of the leg survives. The foot that is being photographed was stuck out parallel to the surface of the water so that we could see its reflection in the surface.

in a straight line, so we add dotted lines showing what the brain would interpret:



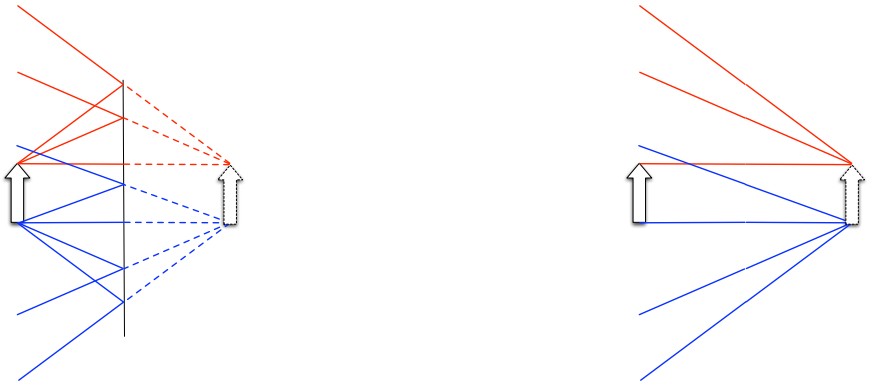
Note that these rays are crossing all over the place! The rays *don't* come back to one place, and the more rays we add the more we see that the light appears to be coming from a whole area under the water. Why when we look straight down on an object does our brain still see the object (fairly) clearly?

The answer is that our brain can only interpret the light rays that actually hit our eyes. Placing a pair of eyes on the diagram as well, and concentrating only on the rays that make it into our eye tells a very different story.



If we look straight down on the object then it appears to be closer than it really is. You can see this quite dramatically be standing knee-deep in a swimming pool. All points from your foot to your knee appear closer to the surface than they really are to someone outside the swimming pool, while the rest of you appears normal. This leads to you looking very disproportionate!

Mirrors can be used to form images as well. This time we will deal with an extended object, so we can tell what the mirror does to the object's orientation. Consider an arrow being reflected in a flat mirror. We draw some lightrays from the head of the arrow and *use the law of reflection to see where they go*. We do the same for the bottom of the arrow. We notice that if we trace the reflected rays back, all the lightrays from the top seem to be coming from the same place behind the mirror (this tracing is shown with a dotted line in the figure on the left). We have formed an *image* of the top of the arrow. We see all the light from the bottom of the arrow seems to come from a different point.

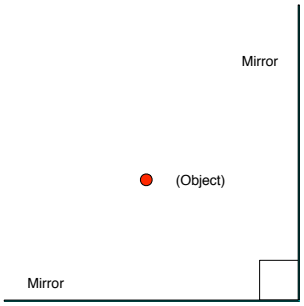


You should check that the rays shown on the left hand side obey the law of reflection.

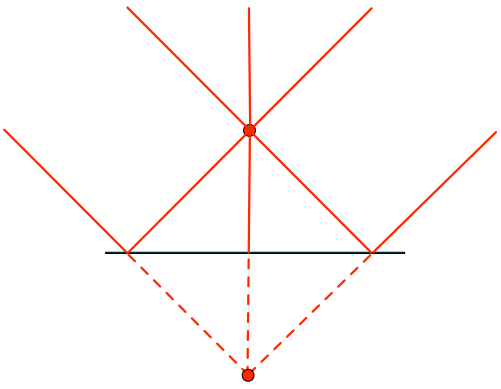
What does this mean? Just like the case of an object immersed in water, the light rays that bounce off the mirror are spreading out, but when our brain traces them back they all seem to be coming from the same place – we have formed an image! One way of thinking about this is that our brain could not distinguish between seeing an arrow and its reflection, or there being two arrows. The light that reaches our eyes would be exactly the same in both cases[4]! Compare the outgoing light rays with the mirror and the outgoing light rays if there was no mirror but simply a second arrow placed where the image of the arrow is. We see that the light rays are the same. (We should also be drawing lightrays that will enter our eyes from the arrow on the far left, but have chosen not to in order to keep the picture less cluttered.)

A complex example: the corner mirror

To emphasise this further, let us consider a slightly trickier example. We will consider a point object between two mirrors at right angles as shown below. What would we see?



To simplify this problem, let us consider only the lower mirror first. Doing the ray tracing we find the location of the image for the lower mirror.



[4]This is making the assumption that there is nothing else around to see in the mirror. Unless you have an identical twin, your appearance in the mirror is a give-away that there are not really two arrows here.

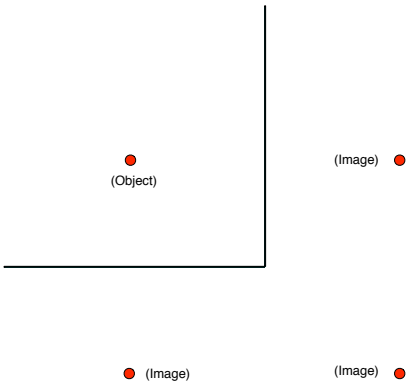
Now here comes the trick: the light rays that bounce off the mirror look *exactly* like they have come from the image. Insofar as the vertical mirror is concerned the light rays that hit it are *exactly* the same as if we had two objects! We are going to pretend that this is really the case – we shall call our original object #1 and the image in the lower mirror object #2. To stop the ray diagrams from confusing us, we shall do the ray tracing for object #1 on the left-hand side, and then do the ray tracing for object #2 on the right-hand side.

Are we done? We have to consider the possibility that the images on the right-hand side themselves have images in the lower mirror.[5] "Image from #2" is not a problem – all the reflected rays head away from the location of the horizontal mirror so it cannot have an extra image. "Image from #1" is more problematic as one of the rays shown in the left hand figure will hit the horizontal mirror.

What we should do is imagine that the horizontal mirror is in place, the vertical mirror is missing and that "Image from #1" is really an object. Note this is what we would have done anyway if we had started with the vertical mirror rather than the horizontal one. Doing this we find that the image of "Image from #1" is in the same place as "Image from #2", so we don't get any new images out.

Now we have convinced ourselves we are done, we can draw the final solution:

[5]We certainly know this *can* happen, as any child that has stood between parallel mirrors in a bathroom and wondered at the infinite number of images of themselves can attest!



When we look at this corner mirror we would think we saw *four* objects – the original object and the three images. While this example was a little involved, it was designed to show that insofar as the outgoing light rays are concerned an image is *indistinguishable* from having an object at that location.

In all the examples so far the light rays have appeared to come from the images, but this has been because our brain assumes that light travels in straight lines. The actual light rays (the solid lines) do not actually meet at the location of the image. Only the *projection* (i.e. assumption that the rays travel in straight lines) of the light rays, shown in the figure as dashed lines, actually cross at the location of the image. These sorts of images are known as *virtual images*. The other sort of image that can exist are *real images*, where the light rays *do* actually cross at the location of the image. We will come across examples of real images when we discuss lenses. (We note that lenses are not the only examples that can give real images; curved mirrors are also capable of giving real images.)

Before starting on lenses, let us summarise the ideas of rays and images:

- Each point of an object (luminous or reflecting) can be taken as a source of diverging rays.
- Light rays within a particular medium travel in straight lines if there is no scattering. Light rays can bend due to medium changes by either reflection or refraction. We will neglect scattering in our discussion.
- Light rays do *not* represent the amount of energy in a wave. We can choose which light rays to draw, depending on which ones are relevant for the problem we are dealing with. In the case of lenses and mirrors

we typically choose to only draw the light rays that hit the lens or mirror.

- If the rays *appear* to be coming from a location, the place the rays track back to is called an *image*.

- To find the complete image of an object, we choose points on the object and see if they form an image anywhere. By doing this for enough points we can reconstruct the image of the entire object.

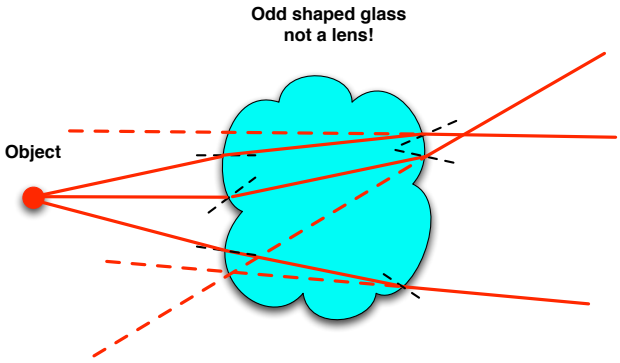
- Reflection and refraction are laws that apply to *all* waves, not just light. Likewise, "images" can be formed for all types of waves as well.

- When dealing with extended objects (like the arrow above) in addition to asking *where* the image is, we can ask what *size* the image is and whether it is "right-way-up" or "upside-down". We will illustrate these concepts with *lenses* in the next section.

8-3-3 Lenses

We have shown that a mirror can produce images by using nothing more than the law of reflection. We have also shown how looking directly into a pool of water also produces an image by using Snell's laws for those rays which enter your eyes. Now we are going to introduce lenses, which are specially shaped materials designed to produce good images by using refraction.

Before getting started, it is worth pointing out something that may seem obvious: lenses are manufactured. A typical transparent object will not be the correct shape to produce a good image. For example, consider the glass blob below. Someone standing to the right would not see the light rays appearing to come from any particular location – they would not get a (clear) image of the object on the other side. We have only chosen to draw three rays here (and show the normal lines used as black dashed lines) but we see that the lines do not appear to be coming from a single location. We would not get a clear picture of the original red object by looking through the blob.



In contrast to our typical blob which refracts light, a lens forms a sharp image. The typical way that a lens does this is by changing its curvature (and hence changing its normal) continuously to ensure that the light rays are actually focused. This may sound contrived, after all how often do we find that the curvature changes in *just* the right way to keep the light in a sharp focus? Well, it **is** contrived! However lenses are incredibly useful in correcting eyesight, magnifying objects and seeing distant stars. It is precisely because these systems are contrived that we must pay so much for glasses – the lenses must be carefully made!

The lenses we will be discussing are either converging (the left picture of figure 8-3.4) or diverging (the right picture of figure 8-3.4). The light bends as it enters the lens (i.e. the light goes from air to the lens material) and as it exits the lens (the light goes from the lens material to air). We are going to simplify the treatment of lenses by treating the light as if it bends only once in the center – a good approximation if the lens is thin. If we knew the precise shape of the lens, we could figure out the normal at every location and use Snell's law ray-by-ray to find how the light exits the lens. Instead we are going to use the fact that the lenses are specially made and have special points called *focal points* to find where light rays go. If you like, it is the job of the person who makes the lens to ensure that the light behaves the way that you want![6]

Before addressing the most noticeable feature of figure 8-3.4 it is convenient to make a definition. The *optical axis* of a lens is the line that goes through the central part of the lens, and is parallel with the normal at the center of

[6]However, by knowing that the material that makes up the lens is glass we can tell that the picture on the left must be a converging lens and the picture on the right must be a diverging lens if these are lenses at all. The precision required of the lensmaker does not let us off the hook for having *some* understanding of why a lens works the way it does.

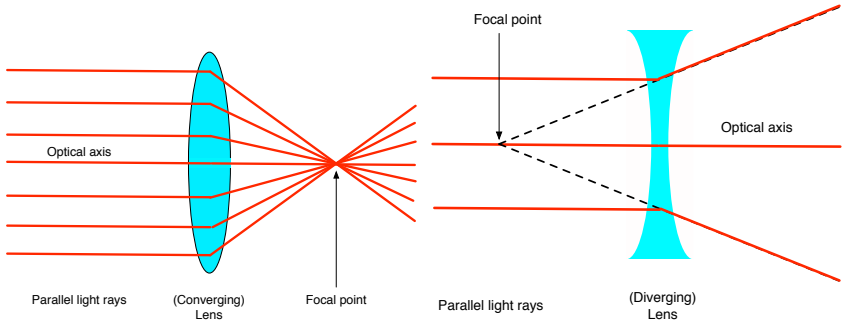
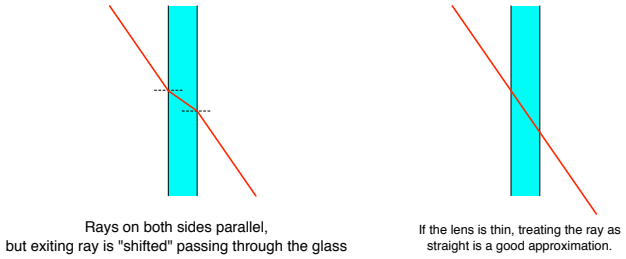


Figure 8-3.4: Showing what happens to parallel light rays as they enter a converging lens (on the left) or a diverging lens (on the right). This diagram also serves to define what a *focal point* is.

the lens. All the rays on the left hand side of the lenses in figure 8-3.4 are not only parallel to each other, but are also parallel to the optical axis.

By looking at the lenses in figure 8-3.4, we notice that each lens has a "special point". For the converging lens the special point is where all the light rays cross, while for the diverging lens this special point is where the light rays *appear* to be diverging from. In both cases the special point is referred to as the *focal point* because it is where the rays parallel to the optical axis are focussed. For a converging lens the light rays actually cross and in analogy with images we call the focus *real*. In contrast no light rays actually cross for the diverging lens. Instead it is our brain's insistence that light travels in straight lines that makes it appear to us that there is a point source behind the diverging lens. In this case the focus is *virtual*.

Before showing how to use the focal point to locate images, we should point out one other important fact: in the middle of the lens the left and right sides of the lens are parallel. Drawing the ray diagram for a ray that passes from air into a block of glass we see that the ray that exits is *parallel* to the ray that entered.

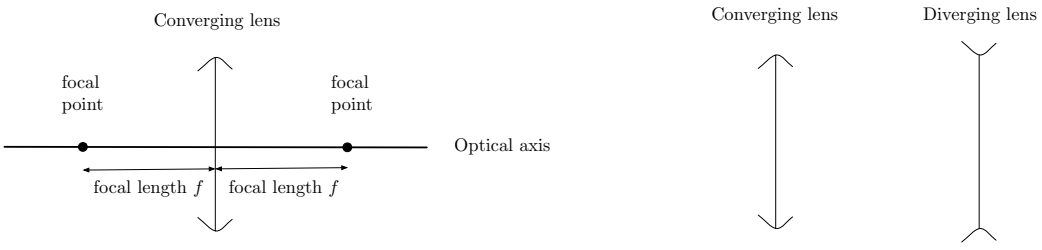


For the *thin* lenses that we are discussing here, the "shift" of the central ray is negligible. We shall treat the ray that passes through the center of the lens as being, to a good approximation, unaffected by the presence of the lens.

8-3-3-1 Ray tracings: using the principal rays

In this section we will learn how we can use the fact that lenses have focal points to allow us to figure out where the rays go, instead of having to apply Snell's law twice for each ray. For each of our lenses there are three rays which are easy to find using knowledge of just the focal point and the position of the object – we call these special rays *principal rays*. We will show how to find the principal rays for both a converging and diverging lens.

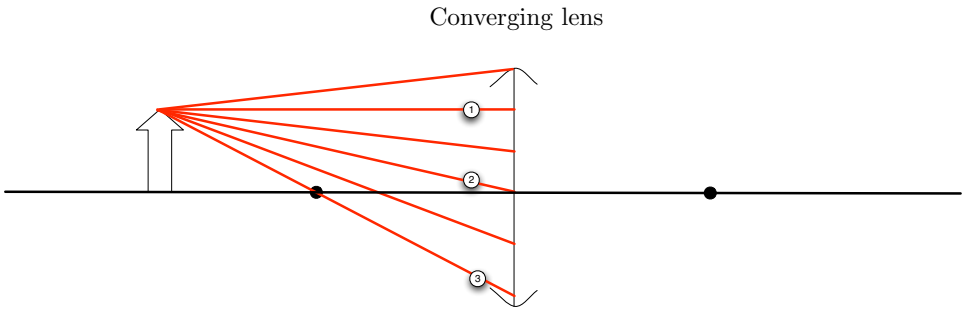
Let us start with some general comments that apply to both converging and diverging lenses. In a lens problem we start with an *object* from which the light comes off. Typically we denote this object as an arrow, so that we can tell by looking at the final image if the object is inverted by the lens. We have discussed what the focal point is for a lens; for a symmetric lens we have two focal points on either side – one where the rays parallel to the optical axis coming in from the left would be focussed, the other where rays parallel to the optical axis coming in from the right would be focussed. The *focal length* of a lens tells us how far from the lens the focal points are. The magnitude of the focal length is the distance from the lens to each focal point, while the sign tells us if the lens is converging or diverging. For a converging lens we take $f > 0$, while for a diverging lens we take $f < 0$. A diagram of a lens and its characteristics (i.e. focal points and optical axis) are shown below.



Finally, as we can see from the picture above, we don't draw out the lens. We are making the approximation that the lens is thin and that refraction (bending) only occurs once at the center of the lens. To make this approximation clear, we replace the lens with a vertical line and the "caps" tell us if the lens is converging or diverging.

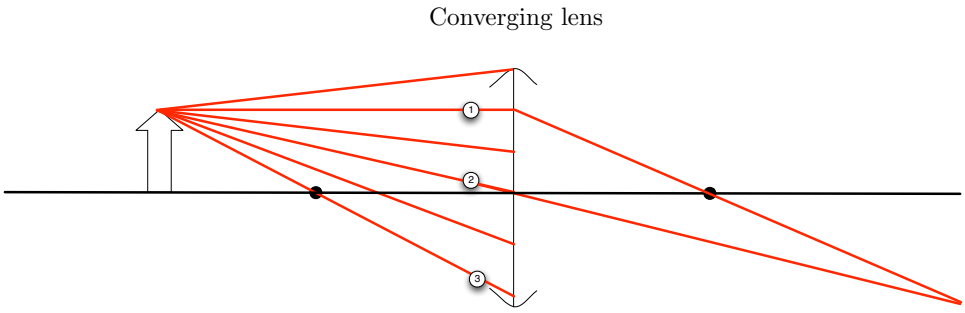
Converging lenses

Let us start by discussing how we would find the image from the object shown below. For the moment let us just ask about the *tip* of the arrow. We know that to find the image of a point we must look at multiple rays from that point and see where they go. We start by drawing many different rays that come off the object



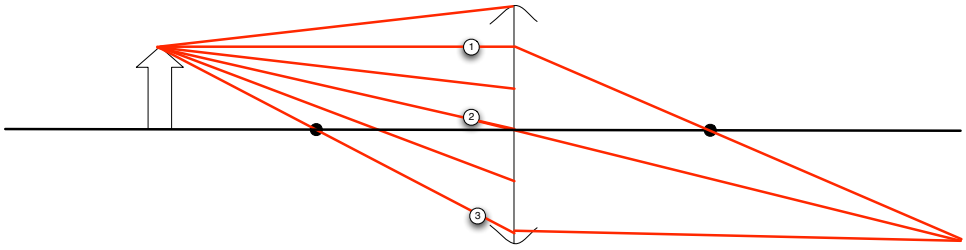
Each of these rays end up being refracted. For those special rays labelled 1, 2 and 3 we can write down where they go straight away.

From our earlier picture, we know that rays that enter the lens *parallel to the optical axis* (such as ray 1) end up going through the focal point on the other side of the lens. We have also discussed how the light that passes through the center does not deviate significantly. For the rays labelled 1 and 2 we can write down what happens to them immediately:



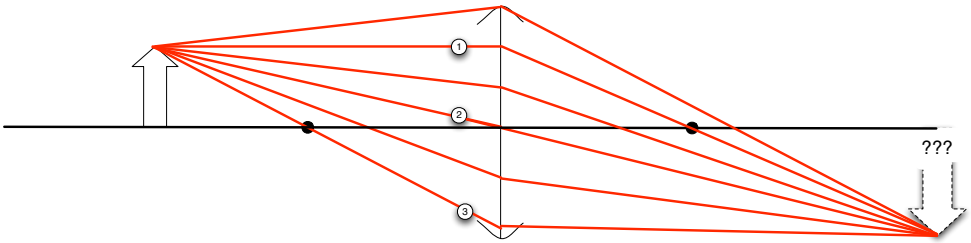
The ray labelled 3 takes a little more thought. We mentioned that refraction was reversible because Snell's law did not care which medium was labelled "1" and which was labelled "2". As a consequence by looking at our converging lens in figure 8-3.4, we see that light that comes *from* the focal point will get bent back parallel. The diagram below shows in detail what happens to ray 3:

Converging lens



Notice that all these rays cross at a particular location. That is where the arrow tip will appear to be for someone who is standing on the far right of the lens. The light rays look like they are leaving this point, just as if the real arrow tip had been placed here. Because the lens designer (we presume) made the lens precisely, all the other light rays should be focussed here as well. Drawing in all the light rays the picture we get is

Converging lens



Notice that we only know where the image of the *tip* of the arrow is going to be from this analysis. Because we don't know when to stop drawing we have simply drawn question marks at the bottom of the arrow. Points on the optical axis are slightly tricky because all three of the principal rays presented here pass through the center of the lens. We will come back to the issue of the optical axis after discussing diverging lenses. Because the light rays actually cross in this example we would call this a *real* image.

If we are only interested in locating the image, then we only need to find out where the three principal rays intersect.[7] The three principal rays are:

1. The ray that is going into the lens parallel to the optical axis; this ray gets bent to go through the focal point (see figure 8-3.4.)

2. The ray that passes through the center; this does not bend.

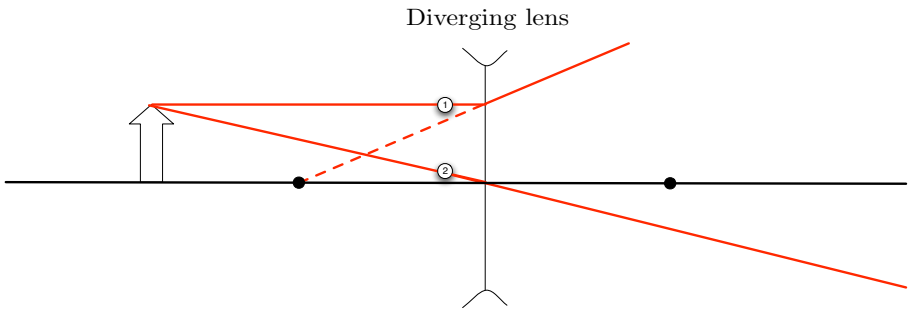
[7]Technically, we only need to find out where *two* of the principal rays intersect. We use three because we can, and it provides a check that we did our other rays correctly.

3. The ray that comes from (or appears to be coming from) the focal point to the lens; this ray gets bent back so it is parallel to the optical axis.

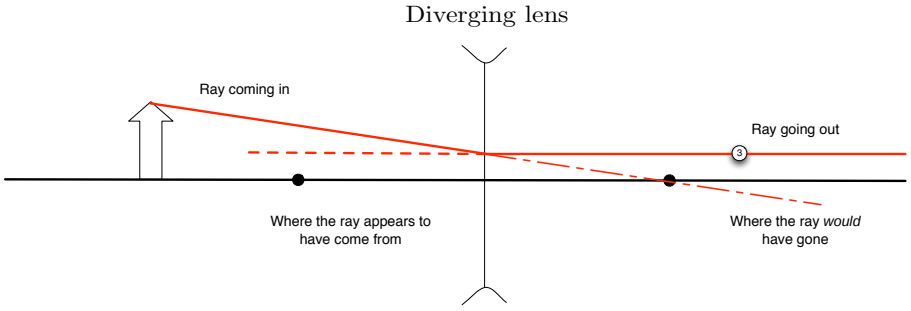
In typical optics problems we will only draw the principal rays, although it is important to realise that all the light rays that pass through the lens and are not obstructed make the final image.

Diverging lenses

Our procedure for diverging lenses is similar to the one presented for converging lenses. We start by referring to how we have defined the focal point in figure 8-3.4 to give us information on how the light should behave. The light which comes off the object parallel to the optical axis gets bent so that the ray appears to be coming *from* the focal point. We label the light ray as ray 1 in the picture below, and draw a dotted line to show where the light ray appears to have come from. We also include ray 2, which passes through the center undeflected.

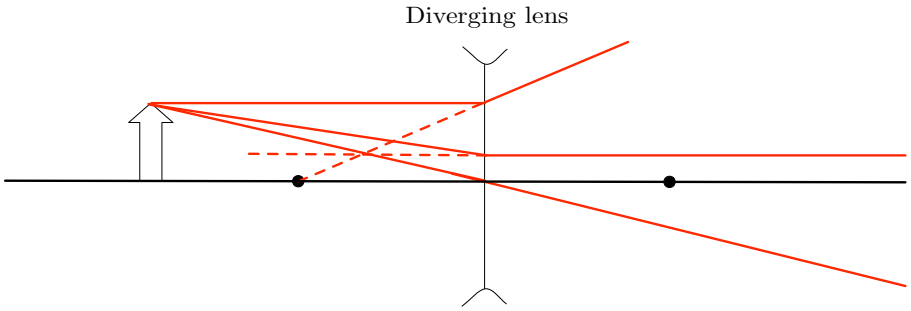


The third principal ray is obtained by considering the reversibility of refraction, just as we did for the converging lens. We know that any light that travels parallel to the optical axis gets bent so that it appears to come *from* the focal point. Thinking about the light travelling the other way, we have that the light that *would* hit the focal point on the other side of the lens would get bent back parallel to the optical axis. We show this below, where we note that the *dotted* line indicates where the light ray *would have gone* had the lens not been there.



Normally we do not draw where the light ray *would* have gone, as is only useful for showing you where to draw the line. The dotted line that shows where the light appears to have come from is incredibly important however, as a person viewing the light from the lens would think that this ray would have travelled horizontally.

So let us put all three principal rays together:



We note that unlike our previous example, the light rays (solid lines) do not actually cross anywhere. However, someone on the right of our lens looking into would think that the light is coming from a smaller arrow located at the point where all the light rays *seem* to be coming from (dashed lines). The dashed lines cross, and our brain interprets this as a small arrow located at the point where all the dashed lines cross even though no actual light rays (solid lines) cross there. We call this a *virtual* image.

To make it slightly clearer why the image seen by a person on the right is the same as a small arrow where the dotted lines cross, it is useful to remember what an image *is*. We could ask "if we had no lens available, what size object would we need, and where would we need to place it so that the light that *reaches us* is exactly the same?" The answers to these questions determine the size and location of the image. For example, if we replace the original arrow and lens with the small arrow located at the point where the

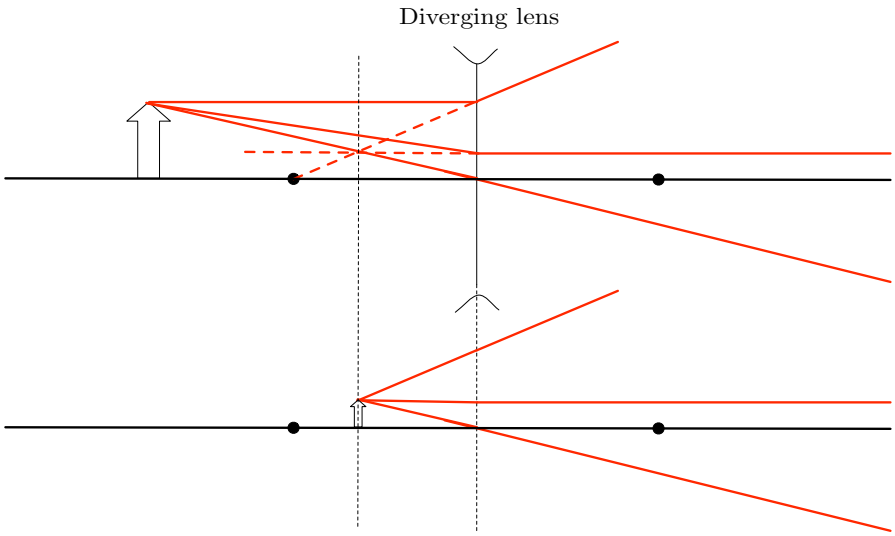


Figure 8-3.5: To the right of the position of the lens in the top diagram, there is difference between the light rays coming off the object and going through the lens **or** the light rays coming directly from the image. Hence our brain cannot distinguish the two scenarios.

dotted lines intersect, we see on the right that there is no difference in the light rays on the far right of figure 8-3.5. If you are standing to the far right the only information your eyes have is the light rays that enter them, and you have no way of distinguishing these two situations. Of course we can see that before the location of the lens the light rays are vastly different, but that does not affect an observer to the right of the lens.

We finish this section by summarising how to find the three principal rays for a diverging lens:

1. The ray that travels parallel to the optical axis gets bent so that it appears to be coming from the focal point on the object's side of the lens.

2. The ray that travels through the center is not deflected.

3. The ray that *would* hit the focal point on the far side of the lens is bent to become parallel to the optical axis.

The optical axis

So far we have found the image location for the tip of the arrows presented. Strictly speaking we should see where the image of the base of the arrow is. This leads to a problem for the examples considered so far because the base of the arrow has been on the optical axis. For both converging and diverging lenses the optical axis is a little tricky. The reason is that the ray that goes through the center of the lens *also* goes through both the focal points. Our normal construction of the three principal rays *fails* for any point on the optical axis as all three principal rays are the same.

What we do know, however, is that points on the optical axis have images on the optical axis. We know this because the one principal ray we do have is always on the principal axis and we know the image is where *all* the rays that pass through the axis meet up (or appear to come from). Since we have one ray that is always on the optical axis the image must also be on the optical axis.

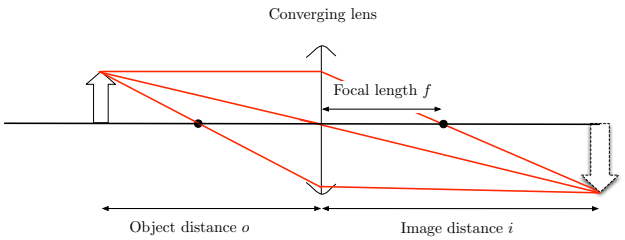
So how do we locate *where* along the optical axis the image forms? One way of finding it is to draw a point that is very close to the optical axis but not quite on it, and perform the ray tracing for that point instead. Another way uses information we will introduce later, the thin lens equation. Whichever way you choose to do it, the result is that the image distance for the base of the arrow is the same as the image distance for the tip of the arrow, provided the tip is directly above or directly below the base. Thus we shall draw our images of arrows down (or up) to the optical axis. If the base of your object is *not* on the optical axis you would be well served to do a separate ray tracing for the base to find where its image is located.

8-3-3-2 The thin lens equation

The problems that we are doing involving lenses have been finding images for a given object and lens combination. We can think of other questions that will be of interest to us later, such as "where would we have to place an object so that the image ends up in a particular location?" or "given this object, and if I want the image to be at a certain location, what is the focal length of the lens I should use?". All of our lens problems, no matter how phrased, come down to relating the object, image and focal length of the lens to one another.

We have already learned how to do ray tracings. Although we have presented ray tracings as a way of going from an object to an image, with some care you can answer any of the questions posed above. The drawback of using a ray tracing is that if we want an accurate answer we must make sure that all our lines are carefully measured and that we setup the problem *to scale*. While a *rough* ray-tracing is often useful for figuring out what sort of answers we should expect, there is an easier way of getting the precise location of an image: using the thin lens equation.

The location of the image along the optical axis depends only on how far away the object is away from the lens, and the focal length of the lens itself. It does not, for example, depend on the height of the image. We introduce o as the distance between the object and the lens, and i as the image distance. The *magnitude* of i is the distance between the lens and the image, but i can be either positive or negative. We choose a convention where $i > 0$ for a real image, and $i < 0$ is for a virtual image. Note that these are definitions for the *sign of* i, **not** the definition of if an object is real or virtual. We remind the reader that we have introduced the concept of a focal length which is positive for converging lenses and negative for diverging lenses. The object distance o is always positive.[8]



$i < 0$	Image is virtual
$i > 0$	Image is real
$f > 0$	Converging lens
$f < 0$	Diverging lens
$o > 0$	(Almost) always

These three quantities o, i and f are related by the *thin lens equation*

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}$$

Looking at our previous ray tracings it is apparent that the image and the object do not have to be the same size. This leads us to define the *magnification* m. We define m as the ratio of the height of the image to the height of the object. Thus, if the magnification is 2 it means that the image is twice as high as the object. We can relate the magnification m to the

[8] It is always a positive number for an actual object. Complications arise when the "object" in question is really an image from another lens in which case it is possible to get a negative object distance. We will neglect this subtlety throughout Physics 7C.

object and image distances i and o via

$$m = -\frac{i}{o}.$$

Notice that the magnification can be negative. If the image is real (so that $i > 0$) then $m < 0$, meaning that the image would be upside-down. A virtual image has $i < 0$ and m is positive, telling us that the image is upright.

The advantage to using these particular conventions (rather than conventions based on which side of the lens that we are discussing) is that we can use the *exact same* conventions when discussing curved mirrors with focal points.

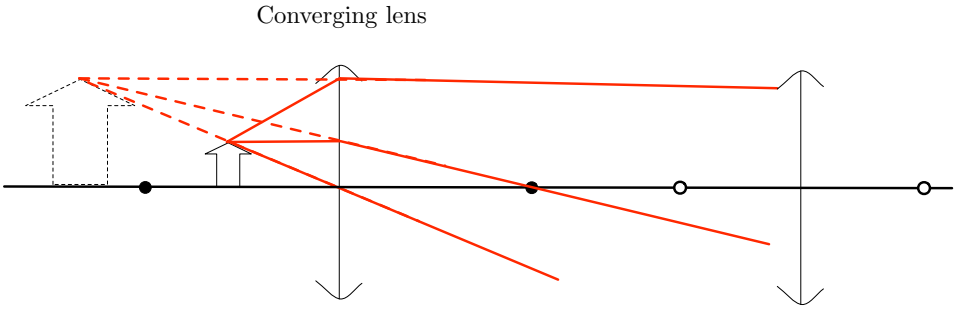
Test yourself:

If an image is bigger in size than the original object, what does this tell us about the magnification? Does it matter if the image is upright or inverted?

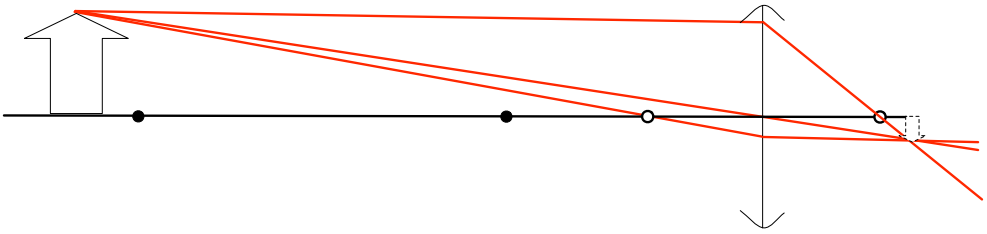
8-3-3-3 Multiple lenses

Once we have single lenses under control, dealing with systems with more than one lens is not much more difficult. We start by looking at only the first lens. The first lens creates an image from the object, and we can find this image location using the techniques we have already discussed. But the whole point of an image is that the light appears to be coming from the image (regardless of the image type). Thus we can replace the object and the first lens by pretending that the first image is itself an object with light coming off it. We then ask what the next lens in our system does to this "new object".

We illustrate this process by way of example. Let us look at a two lens system, and indicate the position of the focal points from the first lens with filled circles, and for contrast indicate the position of the focal points for the second lens with hollow circles. We find the image from the first lens by doing a ray tracing



The combined effect of the object and the *first* lens is to make it look like the light rays are emanating from the image shown as a dashed line. To finish the problem we treat the image as an object in its own right:



8-3-3-4 Applications

Cameras

Here we'll model a camera as a box with film on the back wall, which acts as our screen, where our (real) image should ideally be located at. The lens has a fixed focal length f, and is able to slide in and out of a tube in front of our box. While this is a rather simple model, it is sufficient to explain how most cameras (both film and video) work.

Note that since the camera lens produces a real image, it will appear upside-down on the film negative. This is taken into account in the film developing and printing process, as this negative is used to project yet another real image onto the photograph print. This print is then right-side-up, unless your film developing service loads the negative into their processor incorrectly. If you have a Polaroid camera, which makes a direct print from exposure, the light from the lens must be flipped off of an internal mirror before exposing the Polaroid picture.

Since our simple camera has a lens of a fixed (positive) focal length f, then the lens to image distance i must vary for different object distances o. In fact, if you inspect the thin lens equation, as the object distance o decreases

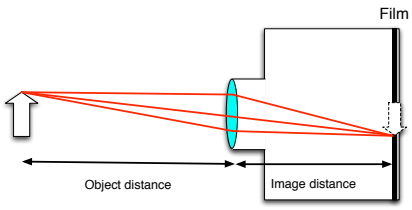


Figure 8-3.6: Cameras change i in order to compensate for varying o

(since f is fixed), then the image distance i must increase. You may have seen this for yourselves, as the lens barrel on your camera must be moved outwards to focus on close-up objects.

Note that "focus-free" cameras are just that – the lens-to-film distance is permanently fixed, such that the images from a select range of object distances will all be at least tolerably focused on the film. Needless to say, photographs from such cameras are not of the highest quality, but these cameras are plentiful because they are cheap and less prone to mechanical failure than are manual focusing and auto-focusing cameras. "Disposable" (though the lenses are recycled) cameras have fixed lens-to-film distances.

Test yourself:

Why can't you make a camera that focuses light onto a film with a diverging lens?

The eye

In contrast to a camera we cannot change the image distance significantly for our eye. That is because the lenses are at the front of the eye (we have both a crystalline lens and the cornea contribute to the bending of rays) and the receptors are on the retina, located at the back of the eye. To get a clear image the image distance i must be the same as the diameter of our eyeball. For most people this distance is roughly 1.71 cm. Because both the eye and a camera require focussing light onto a screen, they both require converging lenses.

So if we cannot change i why can we see things at a variety of distances? Unlike a camera, the focal length of our lens can change. We recall that a

lens works by refraction, and while we cannot change the refractive index of our eye the muscles around the eye, referred to as the *ciliary muscles*, can distort the eye's shape.[9] When the ciliary muscles are relaxed the eyeball is (relatively) flat, and the light rays are not bent much as they pass through the lens. This corresponds to having a large focal distance, because it would take a long distance for light rays parallel to the optical axis to converge to a point. When these muscles squash the eye the lens becomes more round, and the normals change more. This change in normal along the lens corresponds to more bending of the light as it passes through the lens. This "squashed eye" has a smaller focal length. Our ability to change the focal length of our eyes is referred to as *accommodation*.

Let us look at the thin lens equation (and remember that we are holding $i > 0$ to be constant):

$$\frac{1}{f} = \frac{1}{o} + \frac{1}{i}$$

For objects a long distance away the $1/o$ term becomes very small, and the focal length of the eye becomes roughly the same as i, the diameter of the eyeball. As we try and focus on objects that are close the $1/o$ term becomes large. This requires that $1/f$ is also large, or that f is small. That is, our ciliary muscles must squash our eyes a lot in order for us to focus on very close objects. These are the muscles that we feel "strain" when we try and focus on close objects.

There is a limit to how much squashing your eyeball can withstand, or that the ciliary muscles can provide. Consequently there is a shortest focal length $f_{\min}$ that your eyes can have, and a closest object that you can focus clearly on. The nearest distance that you can hold an object while still clearly focussing on it is called your *near point* d_{np}[10]. It is *not* the same as $f_{\min}$, rather they are related by

$$\frac{1}{f_{\min}} = \frac{1}{d_{np}} + \frac{1}{i}$$

[9] You may object that if we are distorting the eye's shape then presumably the lens-retina distance would have to change as well. To a good approximation we can ignore this change, and treat i as constant.

[10] There is a slight problem here – the near point is not actually a point at all! By the definition it is either a collection of points or a distance. Unfortunately this terminology is standard, and here we choose to refer to the near point as the shortest distance that you can focus on.

In practice it is much easier to measure d_{np} than $f_{\min}$, because to measure d_{np} you only need to measure how close you can bring an object to your eye while still being able to focus on it. The nominal value for a middle-aged person for the near point is around 25 cm.

Similarly there is a furthest distance you can focus on when you total relax your eyes. This distance is known as your far point, d_{fp}. For "normal" eyesight the far point is infinity – there is no furthest distance someone can focus on. However, if your eyes cannot relax, or your relaxed focal length is longer than your eyeball's diameter then you will have a far point as well. That is, you will not be able to focus on objects beyond d_{fp}.

Let's now consider three common defects of eyesight. Presbyopia (literally, "elderly eyes") is nothing more than the normal loss of accommodation with advancing age. Children can read books much closer to their face than adults, because their near points are very short and their eyes are able to accommodate quite strongly. This ability decreases with age, so near points for children start to lengthen from as close as 10cm, out to 25cm by middle age (the nominal value for the near point), to even arm's length or longer for older people. Typically, everyone will eventually develop presbyopia. When a presbyopic person's near point is farther than o =25cm, glasses or contacts are prescribed to correct this vision defect.

Farsightedness (or *hyperopia*) literally means only far-away things can be seen. This is because the relaxed lens is too flat, or that the distance between the retina and the lens is too short. A relaxed, unaccommodated farsighted eye cannot focus on distant objects. However, slight accommodation can focus distant objects onto the retina, so farsighted people can see distance objects just fine. As objects get closer, then accommodating more will still focus images onto the retina. For close-up objects, then the eye must strongly accommodate to focus images onto the retina. After a certain point, the eye cannot accommodate any further and near objects remain out of focus.

Typically, children with hyperopic eyes will not have a problem with their vision, because they can strongly accommodate their eyes so they can see objects at any distance. However, as they gradually lose that ability as they grow up, then they will gradually not be able to see close-up objects. When a hyperopic person's near point is farther than $o = 25$cm, then glasses or contacts are prescribed to correct this vision defect. This is somewhat similar to presbyopia, but since accommodation is needed to focus on objects

at all distances in hyperopia, eventually hyperopic people will need glasses (i.e., bifocals, or even trifocals) in order to see all object distances when accommodation is lost.

Nearsightedness (or *myopia*) is literally the ability to only see nearby things. This is because the relaxed lens is too curved, or that the retina to lens distance is too long. Since the ciliary muscles can't "unaccommodate" a lens and flatten it out, there is no way that a myopic eye can see distant objects. As a myopic person's far point is closer than $o = \infty$, then glasses or contacts are prescribed to correct this vision defect. However, while it is still relaxed, a myopic lens is able to focus on midrange objects. Accommodation easily allows the lens to focus on nearby objects.

Eye lenses

When considering corrective lenses, we only need to worry about whether or not a (clear) final image can be made on the back of the retina. The way we go about this is by recalling that the eye itself is a lens, and does not care if it is looking at light that comes off an object directly or light that appears to be coming from the image of some object (as would be the case when we are wearing protective lenses).

This gives us a strategy for modelling corrective lenses. We need to use a corrective lens because the object that we wish to focus on is closer than our near point or further away than our far point. The corrective lens creates an image of the object, and as we learned in our treatment of multiple lenses looking at the object through the corrective lenses is *indistinguishable* from trying to use our uncorrected eyesight to look at the *image* of the corrective lens. Provided the corrective lens places the image between our near point and our far point we will be able to see the object in question. By giving an object range that we wish to see (e.g. all objects up to 10 cm from my face) and knowing the near and far points, we can figure out what the focal length of the corrective lenses is required.

One word of warning: typically when people quote the "strength" of lenses the number quoted is not the focal length. Instead it is the number of optical strength, which has SI units of *diopters* D

$$D = \left(\frac{1\text{ m}}{f}\right)\text{ m}^{-1}.$$

Note this means that the less lens correction needed (which corresponds to less bending, and a *higher* focal length) corresponds to a lower number of diopters. Also note that depending on whether you need converging ($f > 0$) or diverging ($f < 0$) lenses that the prescription for your lens can be either a positive or negative number of diopters.

8-3-4 Summary

1. Become familiar with the idea of wavefronts and rays.

2. Geometric optics is the approximation that rays always travel in straight lines. This approximation is good provided that the wavelength is much smaller than anything it encounters (i.e. we are neglecting *diffraction*). The geometric optics approximation allows us to perform ray-tracings to locate images.

3. When a wave encounters an interface between two media then part of the wave can "bounce-back" (reflect) while the rest can be transmitted into the other media. The transmitted wave bends, a process called *refraction*.

4. The law of reflection states that $\theta_{\text{inc}} = \theta_{\text{ref}}$, with both angles measured from the normal.

5. Rough objects have a rapidly changing normal and as an effect light is reflected in all directions when it hits the surface. This is called *diffuse reflection*.

6. Each non-absorbing material has a refractive index that describes how quickly the wave travels. The higher the refractive index, the slower the wave travels in that medium. For light the refractive index in a medium is defined as $n_{\text{medium}} = c/v_{\text{medium}}$, where v_{medium} is the speed of the wave in the medium and c is the speed of light in vacuum.

7. To find the direction that light bends, we use Snell's law $n_1 \sin\theta_1 = n_2 \sin\theta_2$. Both θ_1 and θ_2 are measured from the normal.

8. If Snell's law cannot be satisfied then none of the wave can be transmitted; instead it is all reflected. This process is called *total internal reflection*. Total internal reflection can only occur when the light is coming from a faster medium and reaches the boundary between media.

9. Our eyes can only track back the rays that reach our eyes, and so if rays appear to be coming from somewhere then our brain thinks there is an object there. If there is no object there, the object our brain thinks it sees is called an *image*.

10. Images come in two types: *real* and *virtual*.

    - A *real image* is where the light rays actually come to a point and then spread out again. This sort of image can be placed on a screen.
    - A *virtual image* is an image where the light rays do *not* cross, but our brain traces back the rays and is tricked into *thinking* that they cross.

11. For thin lenses or particular curved mirrors there is a focal length f. The relationship between the object distance o and image distance i is

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}.$$

    If i is positive, this is a real image, whereas if i is negative this is a virtual image. Lenses work off the principle of refraction; this is *not* a special new law of nature.

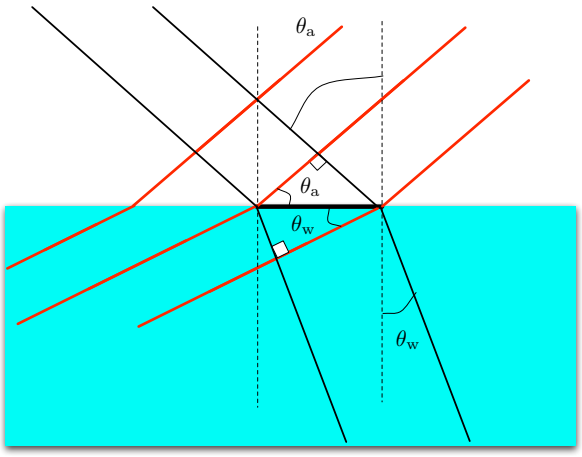
12. The image of an object is typically a different size. We use the *magnification* $m = -i/o$ to describe the change in size; $m = 2$ for example means the image is twice as big as the original object. If the magnification is negative then this means the image is inverted.

8-3-4-1 Derivations*

Three results that we have simply presented have results we could have derived rather than simply stated. We present the derivations here for the interested reader. While not necessary to apply these equations, understanding these derivations will deepen your understanding.

Snell's law

Let us draw both the peaks (as wavefronts) and the rays of light together for light that is travelling from air into water.



Look at the distance between the wavefronts on the boundary, shown as a bold line between the two indicated normal (dashed) lines. Let us call this distance h for hypotenuse, because it is the hypotenuse of both the right angled triangle in the water and the right angled triangle indicated in the air. We know that the distance between the wavefronts (which makes up the opposite side of these triangles) is given by the wavelength in that medium. Writing this out for the triangle in the water we have

$$\sin\theta_w = \frac{\lambda_w}{h} \Rightarrow h = \frac{\lambda_w}{\sin\theta_w}.$$

For the triangle in the air we have a similar relationship:

$$\sin\theta_a = \frac{\lambda_a}{h} \Rightarrow h = \frac{\lambda_a}{\sin\theta_a}.$$

Because we know that the hypotenuse is the *same* in both of these equations, we are lead to the conclusion that

$$\frac{\lambda_w}{\sin\theta_w} = \frac{\lambda_a}{\sin\theta_a}.$$

By multiplying this equation by f and recalling that $v_{\text{wave}} = f\lambda$ we have

$$\frac{f\lambda_w}{\sin\theta_w} = \frac{f\lambda_a}{\sin\theta_a}$$
$$\frac{v_w}{\sin\theta_w} = \frac{v_a}{\sin\theta_a}$$

Finally we recall that $v_{\text{a}} = c/n_{\text{a}}$ (and a similar result for water) we have

$$\frac{c}{n_w\sin\theta_w} = \frac{c}{n_a\sin\theta_a}.$$

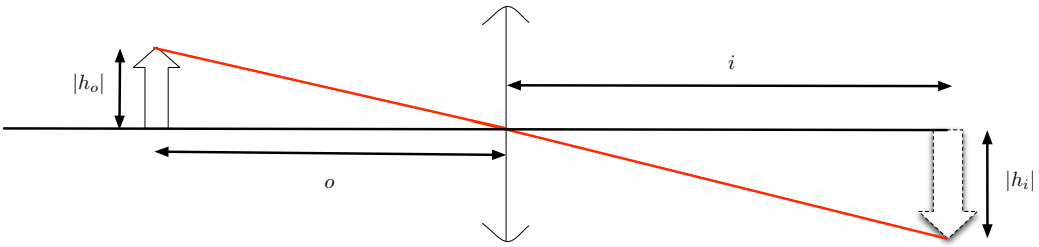
This result only holds if

$$n_w \sin\theta_w = n_a \sin\theta_a,$$

which is precisely Snell's law, is true.

Magnification

Look at the principal ray that goes through the center. Because it is a straight line, the gradient does not change.



We see that the gradient on the left hand side is

$$\text{gradient} = \frac{\Delta y}{\Delta x} = \frac{-h_o}{o}.$$

We can use the information on the right-hand side to calculate the gradient we get

$$\text{gradient} = \frac{\Delta y}{\Delta x} = \frac{h_i}{i}.$$

Because this ray does not bend, we know these gradients are the same. Therefore:

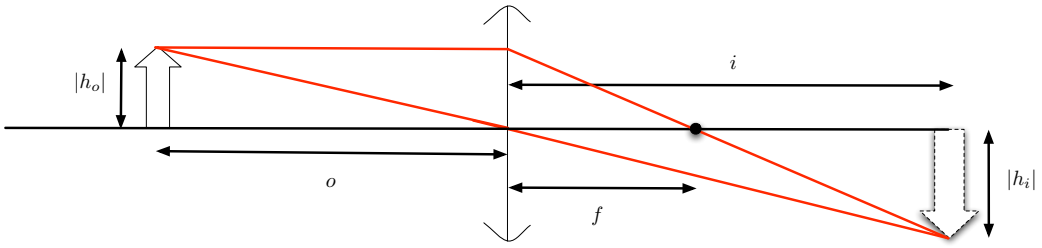
$$-\frac{h_o}{o} = \frac{h_i}{i}$$

Rearranging this equation we have

$$m \equiv \frac{h_i}{h_o} = -\frac{i}{o}$$

The thin lens equation

We will prove the thin lens equation for a converging lens that produces a real image. The other cases can be shown in a similar manner. Unlike the magnification, we shall pay attention to the ray that goes parallel to the optical axis, and goes through the focal point.



We know that there are two ways of calculating the gradient of the ray that passes through the focal point. The first manipulates the fact that the incoming ray has the same height as the object, but drops to the focal length within a focal distance

$$\text{gradient} = \frac{\Delta y}{\Delta x} = -\frac{h_o}{f}$$

The second way of calculating the gradient uses the fact that the height of the ray drops to the location of the image in the distance i. Because $h_i < 0$ we should be slightly careful with the sign of Δy

$$\Delta y = y_f - y_i = h_i - h_o \Rightarrow \text{gradient} = \frac{h_i - h_o}{i}$$

We can get rid of h_i by using the magnification

$$h_i = m h_o = -\frac{i}{o} h_o.$$

Writing out our gradient again we obtain

$$\text{gradient} = \frac{-\frac{i}{o} h_o - h_o}{i} = -\left(\frac{1}{o} + \frac{1}{i}\right) h_o.$$

As a straight line has a constant gradient, the segment we use should not matter. Therefore these two expressions for the gradient must be equal:

$$-\left(\frac{1}{o} + \frac{1}{i}\right) h_o = -\frac{h_o}{f}$$

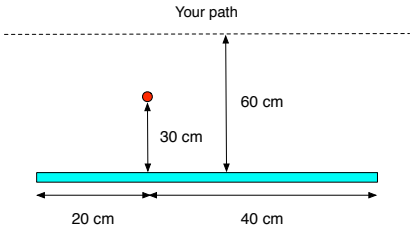
Cancelling the $-h_o$ from both sides leaves the thin lens equation

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}$$

8-3-5 Exercises

1. A ball is sitting 30 cm in front of a plane (i.e. flat) mirror that extends a long way. Draw a ray-tracing and show the location of the image of the ball. Is the image real or vritual?

2. Now the ball is sitting 30 cm in front of a *finite* plane mirror. You are walking by, 60 cm in front of the mirror (shown as the dotted line). How far left of the ball can you stand while on the dotted line and still see the image of the ball? Where is the image located? Is it real or virtual?



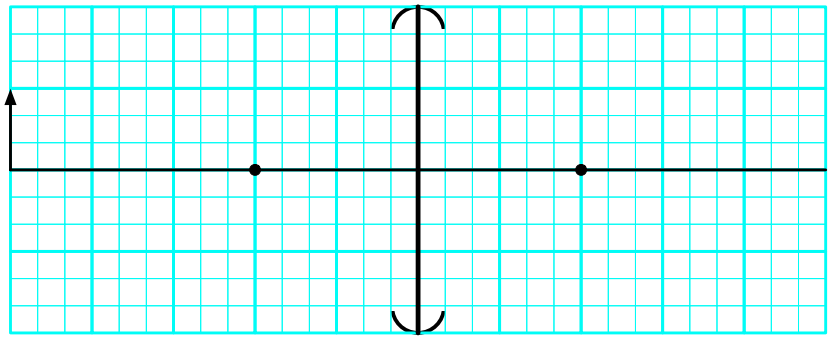
   Note: diagram is *not* to scale.

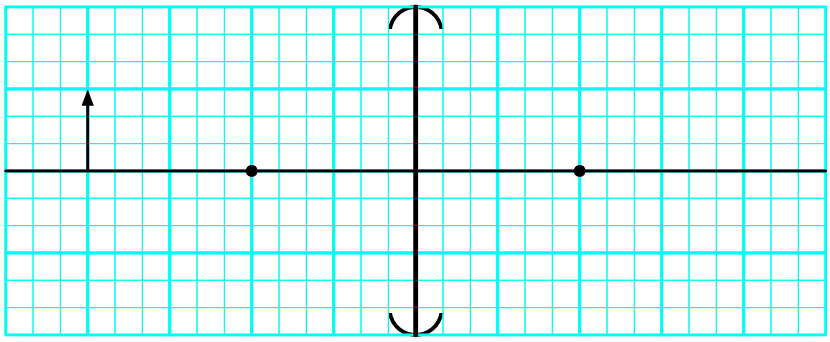
3. For the same scenario as the previous question, how far to the *right* of the ball can you stand while on the dotted line and still see the image of the ball?

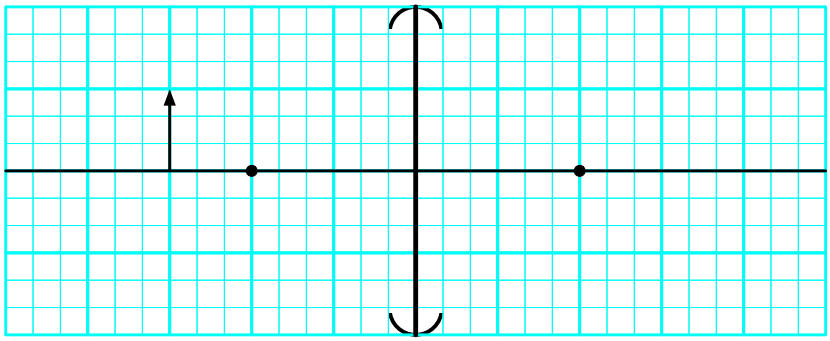
4. On the next two pages are pictures of a lens, and an object indicated by an arrow. The focal points are shown as the black dots. On the diagrams complete the ray tracing and ask yourself the following:

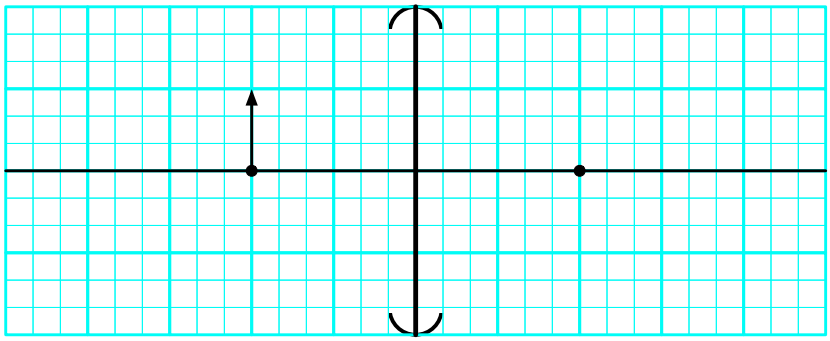
   - Does an image form?
   - Is the image real or virtual?
   - Where is the image formed?
   - Is the image larger or smaller than the original? Is it inverted by the lens? Putting *both* of these together, what is the magnification?

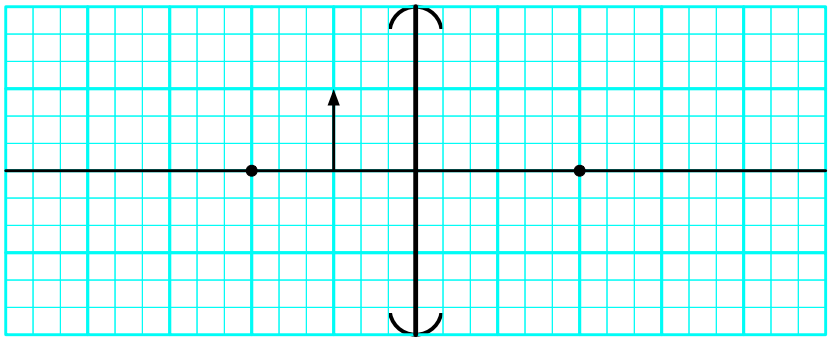
5. Now pick some examples from the 10 rays-tracings, and find the images using the thin lens equation. Is the image where you expect? Is the sign of i the sign you expect? Treat each little square as a distance of 1 cm, so that the lenses have $|f| = 6$ cm.

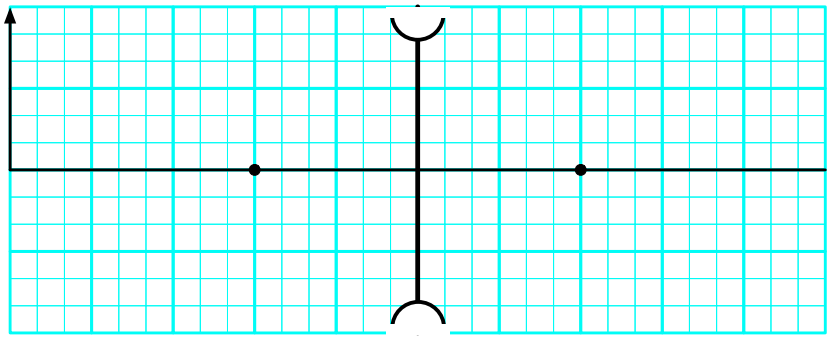


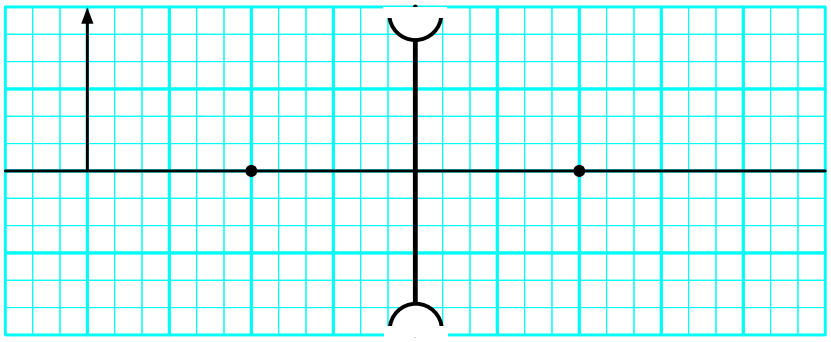


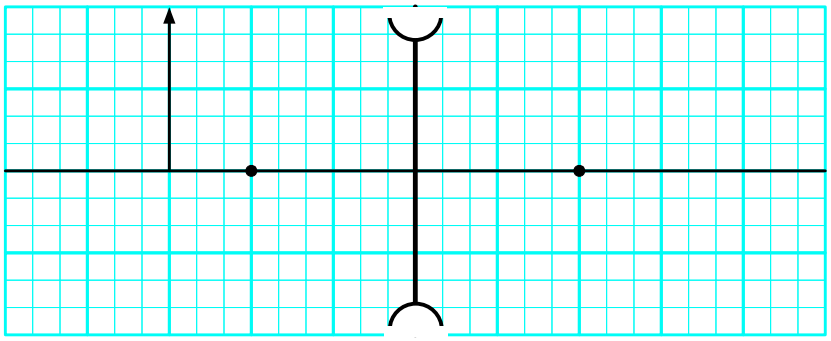


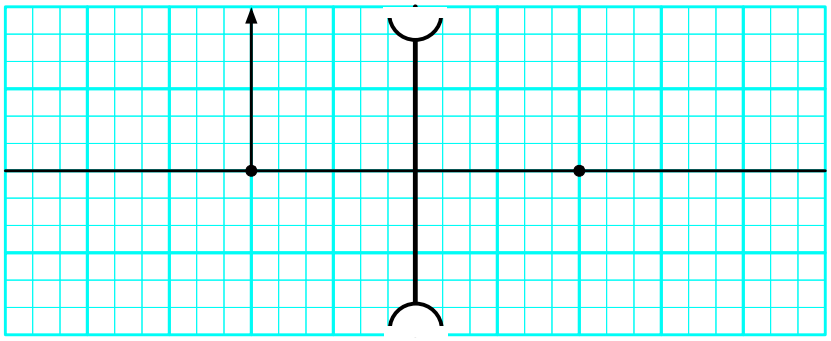


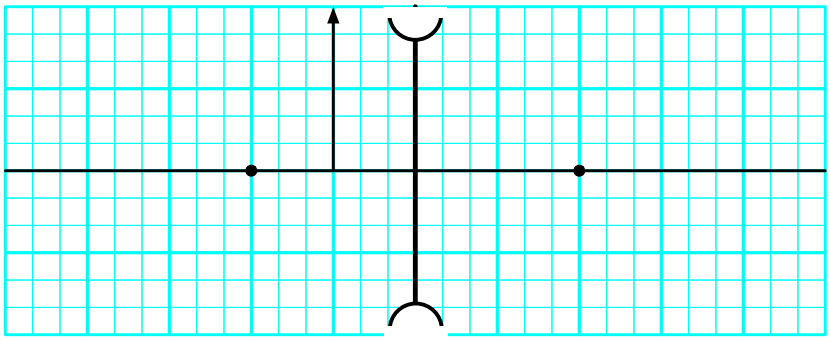












Unit 9

Fields

Unit 9:

9-1: Fields

9-1-1 Overview

Fields are used in almost every part of physics. In physics 7C we will concentrate on the gravitational, electric and magnetic fields. We have some experience dealing with gravity from Physics 7A and 7B, so we will gain our intuition on fields from concentrating on gravity. We will not be able to do anything we could not before after discussing the gravitational field, but it should make discussing less familiar fields easier.

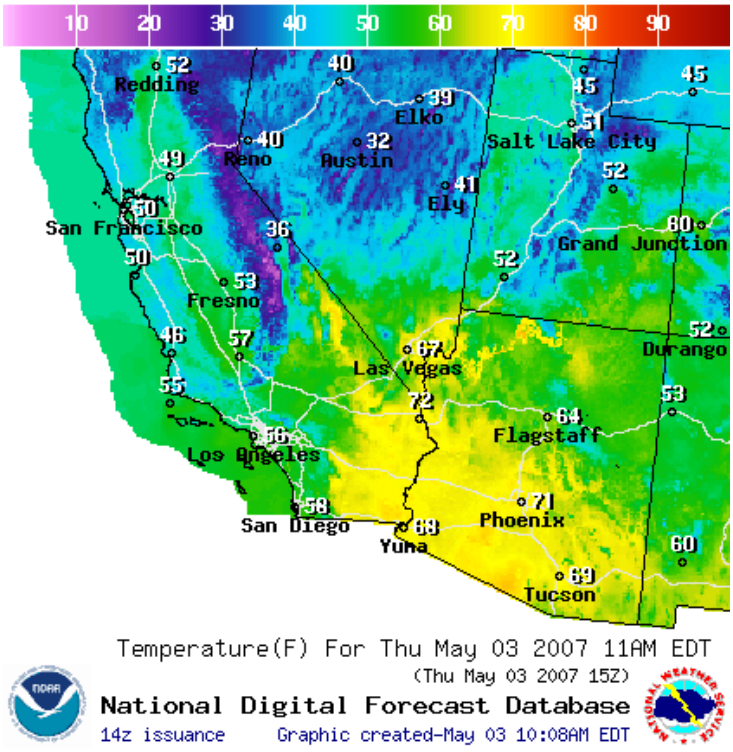
After dealing with gravitational fields we develop electric fields (§9-2) and magnetic fields§9-3. Here the language of fields is much more useful than thinking about objects interacting directly with one another. The last section in this unit (§9-4 shows that electric and magnetic fields are closely related and can propagate – a phenomenon we commonly refer to as "light"! For this problem, fields are indispensable: neglecting them would lead to a violation of both the conservation of energy and conservation of momentum! So when we are learning how to calculate the *same* quantity with or without using fields (as we will do here under the names of *direct method* and *field method*) it is worth keeping in mind that there is a good reason for going through this!

By the end of this chapter you should have familiarised yourself with the notion of field and potential, and be comfortable with how these things are different from force and potential energy. You should also be familiar with three different representations of the gravitational and electric field: the field map representation (§9-1-3-3), the field line representation (§9-1-3-3) and the equipotential representation (§9-1-4-2). The magnetic field is quite different in character, as explained in detail in §9-3. For the magnetic field, only the field map and field line representations are useful.

9-1-2 What are fields?

The idea of a field is that there is some physical quantity that has a value "everywhere", that can either change from location to location or can stay the same. Both fields that vary in space and fields that are constant in (regions of) space are important. A field can also change in time, so any field that we discuss is a function of both position and time. While this is an easy thing to state, it is rather abstract, so let us become more familiar with this definition by looking at some examples:

Temperature field



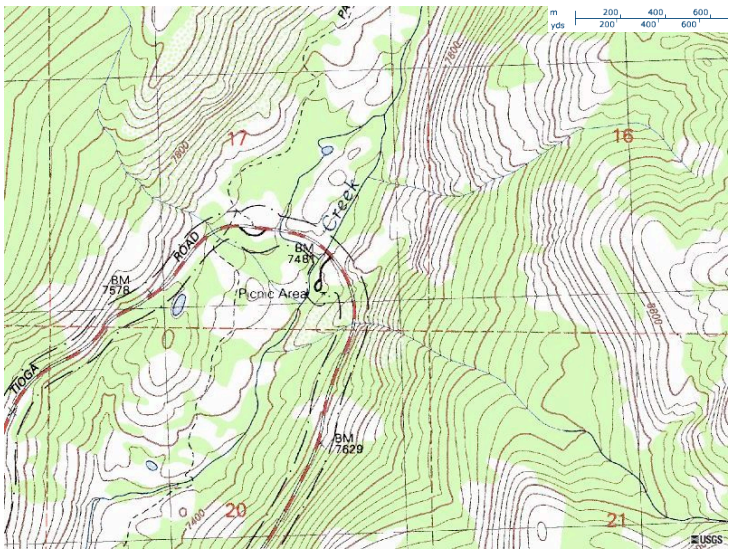
Courtesy of the National Weather Service

The weather map on the news is an example of a field. There is not one universal temperature; the temperature depends on when you ask (which is why the map changes day to day) and where you are asking about. The question "what is the temperature?" does not mean much, as the temperature varies from place to place and changes in time. A question like "what is the temperature in San Francisco now?" can be answered because we have specified both the place and the time.

A couple of points to note about the weather map:

- It shows temperatures at a *particular* time, so what is being shown is $T(x, y, t = \text{today})$. The full field $T(x, y, t = \text{today})$ could be represented by an entire archive of all previous (and future!) temperature maps.
- On this weather map the temperature is given at a number in certain locations, and at places with low population densities is given as a colour.
- On some weather maps the temperature is only shown for selected locations. Even in the places where a temperature is not shown there is a temperature.

Topographical field



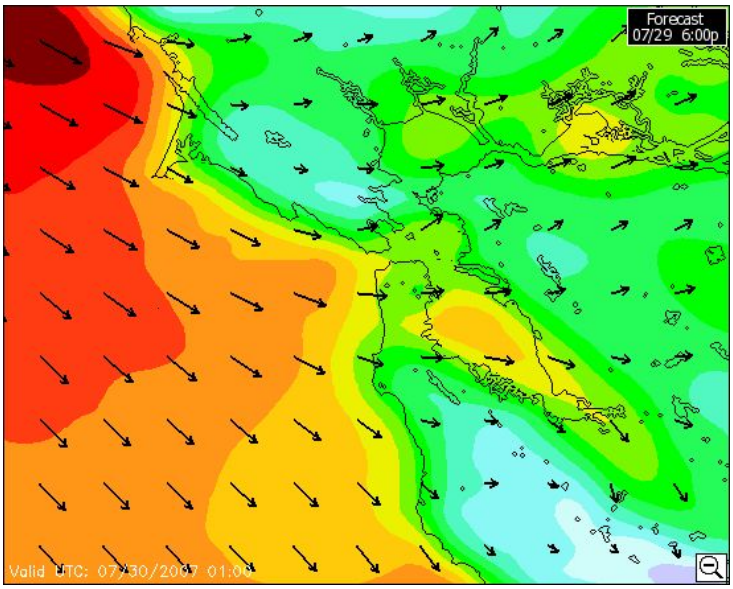
Contour map of Yosemite National Park, courtesy of the National Atlas of the United States

A topography map answers the question "what is the height at this location at this time?". Because the Earth does not shift quickly, we can neglect that it depends on time.

Displacement field

In unit 8 we characterised material waves by looking at $y(x, t)$, which represents the displacement of the wave from its equilibrium position. Thus the displacement depends on both the location being discussed (x) and the time (t) and the wavefunction $y(x, t)$ is a displacement field.

Wind field



Map courtesy of WeatherFlow, Inc.

A wind map shows the velocity of the wind at various locations at a fixed time. This is different from the previous examples because when we look at a particular place and time the field gives us a *vector* – both the wind speed and direction. For this reason, the wind field is a *vector field*. The previous examples gave us a *number* at every point in space, as those were examples of *scalar* fields.

Summary

If a quantity can vary in time and space, then we introduce the idea of a field to reference what its value is at a specific time and location.

In our examples we have also seen two very different types of fields. Our first three examples were *scalar* (number) fields. This is because if we asked, for example, "what is the temperature in Davis at 3:30 p.m.?" the answer that we would get back is a number. The wind field was different, because if we asked "what is the wind doing in Davis at 3:30 p.m.?" the answer we would receive would have both a magnitude (the speed of the wind) and direction. That is, the complete answer to our question would be a vector. For this reason the wind field is referred to as a *vector field*. The fields we will use the most (gravitational, electric and magnetic) are all vector fields.

Warning: definitions can be taken too far

There are also examples that fit into our definition of field that are probably not useful. For example, I could define an elephant field in the following way: give me a specific position and a specific time, and the elephant field will tell you how many elephants were there. While it does fit into our definition, it is not a *useful* thing to do – it is still far easier to talk about individual elephants. So far we cannot use the elephant field to make any new predictions, it is not required by any experiments and does not simplify any discussions.

The point being emphasised here is worrying about whether or not something is a field is not particularly useful. The philosophy embraced here is that we shall consider a quantity to be a field if it is either required by experiment or if it is convenient.

9-1-3 Fields in physics

There are three fields in which we will be interested in for physics 7C: the gravitational field, the electric field and the magnetic field. The most familiar of these is the gravitational field, and so the motivation for using fields will start here.

Let us start by making a simple statement, which is very imprecise: the Earth's gravity is stronger than the Moon's gravity. Justification for this statement comes from watching videos of the astronauts on the Moon and we see that they fall slower and can leap higher. This is adequate for the "person on the street", but in science we must be more precise. What does this statement actually mean? If we calculate the force of the Moon on the Apollo lander, this is much greater than the force of Earth on an apple. Therefore we **cannot** make the blanket statement that the force of gravity on Earth is always greater that the force of gravity on the Moon.

The solution to this "problem" is rather simple: we need to compare apples to apples. If we ask what $\boldsymbol{F}_{\text{Earth on apple}}$ is and what $\boldsymbol{F}_{\text{Moon on apple}}$ is then we find $\boldsymbol{F}_{\text{Earth on apple}} > \boldsymbol{F}_{\text{Moon on apple}}$. More generally, it is true for *any* object X:

$$|\boldsymbol{F}_{\text{Earth on X}}(\text{at surface of Earth})| > |\boldsymbol{F}_{\text{Moon on X}}(\text{at surface of Moon})|\,.$$

This is a more precise version of what we mean when we say that the Earth's gravity is stronger than the Moon's gravity.

We can actually do a little bit better than this. The force of gravity does not distinguish between apples, oranges or skyscrapers. If we could build a skyscraper with the mass of an apple, $\boldsymbol{F}_{\text{Earth on skyscraper}}$ would be the same as $\boldsymbol{F}_{\text{Earth on apple}}$! What this tells us is that to compare the strength of the gravitational field we don't need to use exactly the same object, but only two objects with the same mass.

Now let us tie this to the concept of a gravitational field. Recall from Physics 7B that the force of gravity between two spherical masses is

$$\left|\boldsymbol{F}_{\text{Object 1 on Object 2}}\right| = \frac{GM_1M_2}{r^2} = M_2\left(\frac{GM_1}{r^2}\right) \tag{9-1.1}$$

where r is the centre-to-centre distance, and the direction of the force pulls the masses together. G is known as the universal gravitational constant, and is equal to $6.67\times10^{-11}\ \text{N m}^2\ \text{kg}^{-2}$. Unlike $\boldsymbol{g}$, which is a constant on the surface of the Earth but different elsewhere, G is a *universal constant* meaning that it takes the same value regardless of the problem we are doing. It is because this G is so small that we do not notice the gravitational attraction of objects around us unless one of these objects has an enormous mass.

If we want to talk about "how strong" Earth's gravity is, equation (9-1.1) won't do as it requires a second mass. Let us get around this by asking the question "what *would* the force of the Earth be on an object of mass 1 kg located a distance r away, if it were put there". Now we can calculate this:

$$\left|\boldsymbol{F}_{\text{Earth on 1 kg object}}\right| = (1\,\text{kg}) \times \left(\frac{GM_{\text{Earth}}}{r^2}\right).$$

If we agree to always compare the gravity of an object by referring to what the force *would be* on a second 1 kg mass then we can compare this between different masses. The choice of a 1 kg mass was arbitrary, the only important thing is we always choose the *same* reference mass. The quantity in the brackets, which refers to the Earth and the distance away from it, is the *gravitational field* of the Earth. We denote this $\boldsymbol{g}_{\text{Earth}}$:

$$|\boldsymbol{g}_{\text{Earth}}| \equiv \frac{GM_{\text{Earth}}}{r^2}$$

But not all masses in the world are 1 kg! But once we know $\boldsymbol{g}_{\text{Earth}}$ we can easily calculate the force on any other mass:

$$\left|\boldsymbol{F}_{\text{Earth on object}}\right| = M_{\text{object}}\left(\frac{GM_{\text{Earth}}}{r^2}\right) = M_{\text{object}}\,|\boldsymbol{g}_{\text{Earth}}|$$

This is just $\boldsymbol{F}_{\text{Earth on object}} = M_{\text{object}}\boldsymbol{g}_{\text{Earth}}$. We have seen this relationship many times before (this is why we chose to call the gravitational field $\boldsymbol{g}_{\text{Earth}}$ rather than some other letter). We should keep in mind that $\boldsymbol{g}_{\text{Earth}}$ does depend on r, the distance from the centre of the Earth to the centre of the object.

9-1-3-1 The direct and field model of forces

In the way that we have introduced the gravitational field the field is simply a shortcut. Instead of saying "the force a 1 kg object would feel if placed here due to this source is 5 N" we can simply say "the gravitational field of the source here is 5 N/kg". The field is not necessary to determine the gravitational force between two objects, it is simply convenient. We will see later in §9-4-5 that we actually need to talk about fields if we want energy and momentum to be conserved, but for now we will simply treat them as a shortcut.

With this in mind, we have two separate ways of discussing how a gravitational force acts between two objects. The first is called the *direct method* where we calculate the force by putting numbers into Newton's gravitational law (9-1.1) without any reference to the field:

Direct model:

$$\text{Object \#1} \overset{\text{creates force on}}{\Longrightarrow} \text{Object \# 2}$$

The other way we could think about this is in our new language of fields, which is to think of one mass creating the field and another feeling its effects:

Field model:

$$\text{Object \#1 (“source”)} \overset{\text{creates field}}{\Longrightarrow} \boldsymbol{g}_{\text{obj \#1}} \overset{\text{exerts force on}}{\Longrightarrow} \text{Object \# 2 (“test”)}$$

This is referred to as the *field method*, because instead of thinking of one object directly exerting a force on another we think of one object (referred to as the "source") creating a field and then that field creates a force on the second object (referred to as the "test object"). Of course, as these are calculations of the same thing they both give the same answer. The following example will give an idea of how these two approaches compare:

Example #1:

a) What (gravitational) force does the Earth exert on a 2 kg book sitting on its surface?

b) What gravitational force does the Earth exert on the same book 10,000 km above its surface?

Use both the direct and field methods.
(The mass and radius of the Earth can be found in appendix A)

Solution:

Part a) Direct method

By looking up the mass and radius of the Earth we find

$$\begin{aligned}\boldsymbol{F}_{\text{Earth on book}} &= \frac{GM_{\text{Earth}}M_{\text{book}}}{r_{\text{Earth}}^2} \\ &= \frac{(6.67\times10^{-11}\,\text{N m}^2\,\text{kg}^{-2})(5.98\times10^{24}\,\text{kg})(2\,\text{kg})}{(6380000\,\text{m})^2} \\ &= 19.6\,\text{N}\end{aligned}$$

Part a) Field method

We know that $\boldsymbol{g}_{\text{Earth}} = 9.8$ N/kg at the surface of the Earth. Normally we approximate this to 10 N/kg, but let us be more precise for this example.

$$\boldsymbol{F}_{\text{Earth on book}} = M_{\text{book}}\boldsymbol{g}_{\text{Earth}} = (2\,\text{kg})(9.8\,\text{N/kg}) = 19.6\,\text{N}$$

Notice how the calculation was *much* easier, since we already knew $\boldsymbol{g}_{\text{Earth}}$.

Part b) Direct method

This proceeds almost exactly the same as before. The two tricky points here are that we have to recall that r is the distance from the *center* of the Earth, and to change 10,000 km into metres.

$$\begin{aligned}\boldsymbol{F}_{\text{Earth on book}} &= \frac{GM_{\text{Earth}}M_{\text{book}}}{(r_{\text{Earth}}+10,000\,\text{km})^2} \\ &= \frac{(6.67\times10^{-11}\,\text{N m}^2\,\text{kg}^{-2})(5.98\times10^{24}\,\text{kg})(2\,\text{kg})}{(6,380,000\,\text{m}+10,000,000\,\text{m})^2} \\ &= 3.0\,\text{N}\end{aligned}$$

Part b) Field method

We don't know $\boldsymbol{g}_{\text{Earth}}$ at a distance of 10,000 km from the surface of the

Earth off the top of our head, so we have to calculate it first.

$$\begin{aligned}\boldsymbol{g}_{\text{Earth}}(\text{10,000km above surface}) &= \frac{GM_{\text{Earth}}}{(r_{\text{Earth}} + 10,000\,\text{km})^2} \\ &= \frac{(6.67 \times 10^{-11}\,\text{N m}^2\,\text{kg}^{-2})(5.98 \times 10^{24}\,\text{kg})}{(6,380,000\,\text{m} + 10,000,000\,\text{m})^2} \\ &= 1.5\,\text{N/kg}\end{aligned}$$

Now we can calculate the force the Earth exerts on the book:

$$\boldsymbol{F}_{\text{Earth on book}} = M_{\text{book}}\boldsymbol{g}_{\text{Earth}} = (2\,\text{kg})(1.5\,\text{N/kg}) = 3.0\,\text{N}$$

Because we did not know $\boldsymbol{g}_{\text{Earth}}$ before starting the problem the field method was *longer*. But if we were asked to do the same calculation for a different mass 10,000 km above the Earth's surface we now have $\boldsymbol{g}_{\text{Earth}}$ and could do it much quicker.

9-1-3-2 Which mass creates the field? (Newton's third law)

In the example above, when using the field method we decided that the Earth would create the field and the book would respond to it. This seems quite acceptable, as we are used to the Earth exerting a gravitational force. But what if we decided to use a book and a chair in our example? Which would be the "source" for the gravitational field, and which would be pulled by the field?

To try and work out the answer to this question, let us think about the same problem using the direct model of forces. Calculating the magnitude of the force of the book on the chair gives

$$|\boldsymbol{F}_{\text{Book on chair}}| = \frac{GM_{\text{book}}M_{\text{chair}}}{r^2},$$

where r is the distance between them[1]. Now let us calculate the force of the chair on the book:

$$|\boldsymbol{F}_{\text{Chair on book}}| = \frac{GM_{\text{chair}}M_{\text{book}}}{r^2} = |\boldsymbol{F}_{\text{Book on chair}}|$$

[1] "Consider a spherical chair..." – it turns out the corrections due to these objects not being spherical is unimportant.

These forces have the same magnitude, but pull in opposite directions. This is not a coincidence, but is a consequence of Newton's third law that we learned in 7B:

$$\boldsymbol{F}_{\text{A on B}} = -\boldsymbol{F}_{\text{B on A}}.$$

In the language of the field model we see that the answer is *both* the chair *and* the book create a field. To do the complete problem in the field model we would have to look at

$$\text{Book} \overset{\substack{\text{creates}\\\text{field}}}{\Longrightarrow} \boldsymbol{g}_{\text{Book}} \overset{\substack{\text{exerts}\\\text{force on}}}{\Longrightarrow} \text{Chair}$$

and

$$\text{Chair} \overset{\substack{\text{creates}\\\text{field}}}{\Longrightarrow} \boldsymbol{g}_{\text{Chair}} \overset{\substack{\text{exerts}\\\text{force on}}}{\Longrightarrow} \text{Book}.$$

An important consequence of this is that to be *affected* by a field, an object must also *create* a field of the same type. Note that an object does not feel its own field, only the field of all external objects. But if it feels an external gravitational field, it must also create its own gravitational field to be felt by other objects.

While an object that feels a field must also create the same field, when we are emphasising an objects ability to create a field we refer to it as a *source* object. When we talk about the object responding to an external field, we talk about a *test* object. For gravity, any object with mass is a source object. For the electric field, any object with (electric) charge is a source object. For magnetism, as discussed further in section §9-3 the source is *moving electric charges*.

Test yourself:

In the example Section 9-1, ex. #1 would we have to worry about the force of the book on the Earth? If not, why not?

9-1-3-3 Field lines

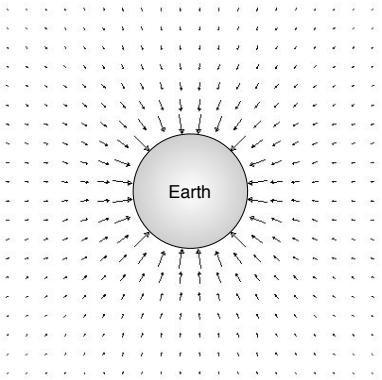
We have learned in section 9-1-3-1 that for point or spherical masses that the gravitational force between them is given

$$\left|\boldsymbol{F}_{\text{spherical mass on 1 kg}}\right| = (1\,\text{kg}) \times \left(\frac{GM_{\text{spherical}}}{r^2}\right) = (1\,\text{kg})|\boldsymbol{g}_{\text{spherical}}|.$$

Therefore we can deduce that the gravitational field created by a spherical or point mass is

$$|\boldsymbol{g}_{\text{spherical}}| = \frac{GM_{\text{spherical}}}{r^2}$$

where r is the distance from the center of the object. The direction of the gravitational field is always pulling inward. Like the wind map, we pick a set of points and draw vectors indicating the direction of the gravitational field:

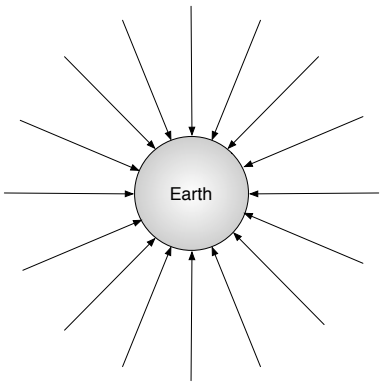


This is called a *field map* of the field. It is the simplest representation of a vector field, as we only have to look at the point in which we are interested to find the strength and direction of the field. Remember that the vectors only refer to the value of the fields at the location that they start, and that the actual length of the vectors is arbitrary – there is an implicit scale to convert an arrow on the page to the correct units[2]. The *ratio* of lengths at two different locations is *not* arbitrary – it tells us about the ratio of the field strength at those two locations.

If we look at the previous field map of the Earth's gravitational field, a long way from the Earth it is almost impossible to read the direction of those arrows. We could enlarge them, but then the arrows near the Earth would have to be enlarged too. If they are enlarged too much they will "poke through" the Earth and become messy. The size of the arrows also means that we get limited information from this picture. To address these shortcomings, we introduce a different representation of a vector field by using *field lines*.

To construct field lines, we draw a continuous lines starting at a point and always going in the direction of the field. An example for the Earth is shown below:

[2]This is not new, this was also the case for the force vectors in force diagrams that you have seen in Physics 7B.



Notice that we can no longer simply "read off" the strength of the field at a particular point as we did on the field map. However this picture contains all the information as the vector map pictures. While we cannot look at the length of the arrows to get the strength of the field, we can instead look at how closely packed the field lines are. The closer they are (by the Earth), the stronger the field. The further apart the field lines are, the weaker the field.

If we start with the field line diagram, we can construct the field at any given point the following way:

- Direction: Take a tangent to the field line at that point. This is the direction the field will be going.
- Magnitude: Given by the "density" of the surrounding field lines.

Vector map

Pros	Cons
Can read off direction Can read off magnitude	Hard scaling issues and readability issues

Field lines

Pros	Cons
Scales well for forces that differ in magnitude. Direction easy to read	Must work to reconstruct the magnitude of the field.

While the vector map is the most direct representation of the field, we will frequently prefer to use the field lines.

Gauss's law

One thing that has not been made explicit in the discussion of field lines so far is that they cannot just start or stop in any location. For gravity, all the field lines start a very long distance away and can only stop when they encounter a mass. If there are no masses in a particular region then the field lines cannot be created or destroyed; they simply keep going. The number of field lines that stop is proportional to the mass of the object encountered. Thus not all the field lines will "stop" because they hit the satellites that

orbit the Earth, although a few will. Don't worry about field lines passing through physical objects – remember that they are only a *representation* of the field.

For the electric field, field lines can start on positive charges and end on negative charges. This makes electric cases slightly more subtle, because if the number of field lines entering is the same as the number leaving it could be that the region has no charges in it *or* it has an equal number of positive and negative charges. For electric fields looking at the number of field lines entering and exiting a region only tells us about the *net charge* in the region, not how many individual charges are in that region.

In either case, in region with no mass (for gravity) or net charge (for electric field lines) the field lines cannot be created or destroyed. Therefore if the number of field lines entering a region is different from the number leaving we must have a source (i.e. a mass or charge) in that region. By knowing the difference, we can figure out exactly how much mass (for gravity) or net charge (for electric fields) is in the region. This is the essence of Gauss's law, which we now make more precise.

Think of an imaginary closed surface, meaning that it has an inside and an outside. Here the term "inside" does *not* have its colloquial meaning; a shape that has an "inside" in this sense means that you cannot get out unless you go through the surface. A box with a lid is a closed surface, but a glass is not as there is a hole at the top (where we tip liquids in or out). Even though I have used physical objects for my examples, we can use any shape our imagination desires provided that it is a closed surface.

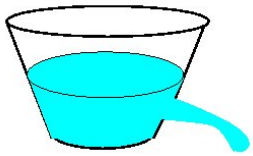
Not closed Closed

Once we have our closed surface, we can ask how many field lines enter it, and how many leave. As it is closed, the only way a field line can get in (or out) is by going through this surface. If a different number of field lines come in than go out, we know that field lines are being created or destroyed inside. That is, there is a mass inside (if we are looking at gravitational field lines) or a net charge inside (if we are looking at electric field lines). If there is no

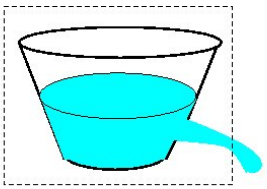
difference between the field lines entering and leaving there is no mass inside or no *net* charge, as overall no field lines are being created or destroyed[3]. The number of field lines going through the surface is referred to as the *flux*. Gauss's law is simply the statement that

$$\begin{aligned}\text{Net flux} &\propto (\#\text{ field lines entering}) - (\#\text{field lines leaving}) \\ &\propto \begin{cases}\text{total mass inside surface} & \text{(gravity)} \\ \text{total charge inside surface} & \text{(electric)}\end{cases}\end{aligned}$$

A useful analogy to a closed surface and field lines is a leaky bucket filled with water. The bucket is not closed, but you can *imagine* a surface consisting of the actual bucket and a lid. If water is leaking out of the bucket, it is also leaking out from the inside of this imaginary surface. Therefore the water must be passing through the surface. As the water cannot pass through the bucket, this must mean the bucket has a hole in it!



Water leaves the leaky bucket



Water leaves the imaginary surface

Of course, you did not need to go through all this to figure out your bucket has a hole in it. But this familiar example is exactly what we do with fields – if the field lines are coming out from somewhere then something is creating them. If we are losing field lines in a region, something is destroying them.

Gauss's law and magnetism?*

So far Gauss's law has been discussed for gravitational fields and electric fields, but no mention of magnetic fields has been made. That is because, to the best of our knowledge, there are no "magnetic charges". Instead all known magnetic fields are created by moving electric fields, and as a consequence magnetic fields do not "start" or "end" but instead make circles. Because magnetic field lines never start or end the number of magnetic field lines entering a surface is always equal to the number of magnetic field lines leaving that surface.

[3]In the case of the electric field, we cannot quite say this. Instead, we should say that the number of field lines being destroyed (by hitting a negative charge) are equal to the number being created. The amount of positive and negative charge must be the same, so there is no *net* charge.

9-1-4 Potentials and equipotentials

9-1-4-1 Gravitational potential

In Physics 7A, we tied together the idea of potential energy and force. We learned that the magnitude of the force was given by the slope of a PE vs. r graph. Just like a force is between two objects, the potential energy is always an energy between two objects.

What we would like to do in this unit is to talk about energy (or something like it), but only for *one* object. We got around this same problem with forces by introducing a new concept – fields – which answered the question "what *would* the force on a 1 kg object if it were placed a distance r away from the source?". What to do with potential energy is now obvious; we should invent a new concept that essentially answers the question "what *would* the potential energy be for a 1 kg object placed a distance r away from the source?". The name for this new concept is *potential* and it is represented by U. The name is an unfortunate choice because while it is closely related to the potential energy they are not the same thing. The relationship between the two is

$$\mathrm{PE}_{\text{grav between obj 1 and obj 2}} = M_1 U_2$$

This equation reminds us that it does not matter which object is considered the "source", although in cases where one object is much bigger than the other it is conventional to treat the larger object as the source.

For a point mass, or a spherical mass of uniform density, the equation for the potential is relatively straight-forward:

$$U_{\text{grav}} = -\frac{GM}{r} + U_0 \tag{9-1.2}$$

where r is the distance from the centre of the mass creating the field. Here U_0 is some arbitrary value. In Physics 7A we learned that the potential energy could not be measured; it was the *changes* in potential energy that could be observed. Absolute potential cannot be measured either, instead only *changes* in potential are observable. If we add the same constant to the potential everywhere, there is no experiment that can tell the difference.

Notice that far away from the mass we have $U_{\text{grav}} \approx U_0$. We will find it convenient to adopt the *convention* that the gravitational potential goes to

zero a long way from the source. This corresponds to choosing $U_0 = 0$, in which case the potential for our point or spherical mass is

$$U_{\text{grav}} = -\frac{GM}{r}$$

As we get further in, the potential is *negative*, and because mass is always positive this tells us that the gravitational potential energy is *negative*.

Let us compare the gravitational potential energy (with zero at infinity) with the Lennard-Jones potential energy you looked at in 7A. There the PE went to zero as r became large (again, by convention), and because the potential energy was negative the total energy could be negative as well. If the total energy of a system was negative this indicated the system was bound, as the kinetic energy cannot be negative. Similar reasoning applies in the gravitational case.

Test yourself:

Starting with (9-1.2), can you get the formula for the potential energy between two masses? Can you go from there to calculate the force between two objects?
(See the diagrams in section 9-1-6 for help).

Test yourself:

Is the total mechanical energy of the Earth (i.e. KE + PE) positive, negative or zero? How can you tell? For this question, take PE = 0 to be a very long way out of the solar system. (Hint: Remember that the Earth is orbiting the sun)

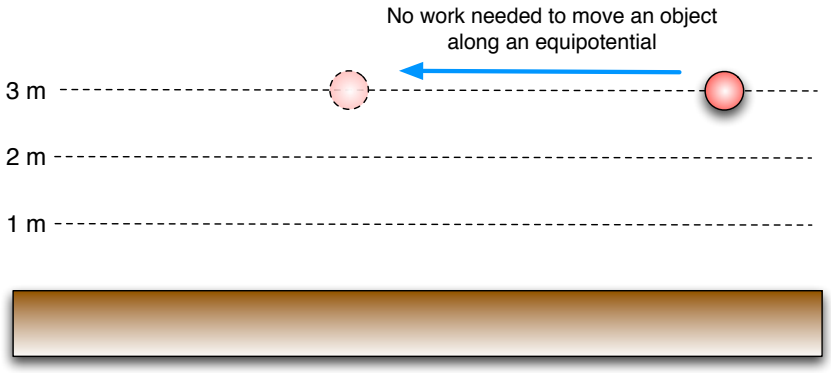


Figure 9-1.1: The height above the ground is the most familiar example of equipotentials (shown as dashed lines).

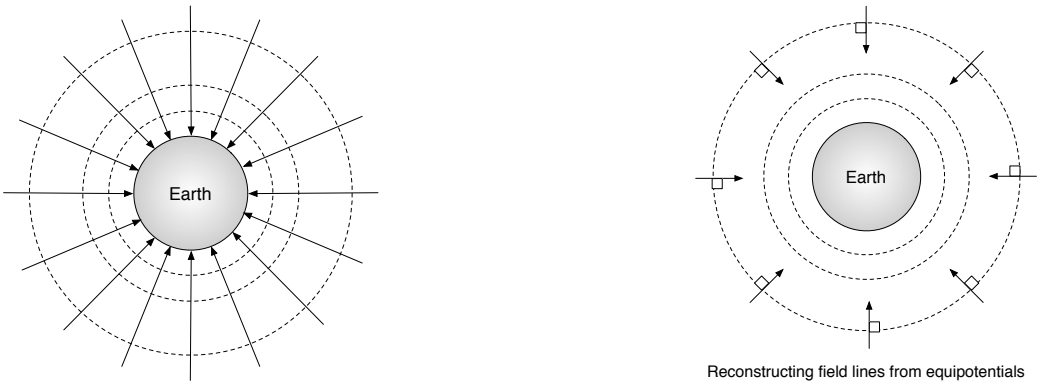
9-1-4-2 Equipotentials

An *equipotential* (i.e. "equal potential") is the continuous curve along which every point is at the same potential. As a consequence, it takes no work to move along an equipotential. From this we can conclude that the force has no component in the direction of motion. The most familiar example of equipotentials are the height above the ground, as shown in figure 9-1.1. We know that the mass only gains gravitational potential energy if the height changes, but moving it horizontally (i.e. along an equipotential) does not change the gravitational potential energy.

For the electric and the gravitational field, the force is always in the direction (or against the direction, for negative charges in an electric field) of the field lines. An equipotential cannot move with or against the field, as this would mean an object would gain or lose potential energy. *All equipotentials are at 90° to the field lines, and any given equipotential only intersects a given field line once.* If an equipotential intersected a field line twice that would mean it was possible to move with (or against) the field and not change the potential energy of an object which is impossible.

We can use the fact that equipotentials and field lines are perpendicular to reconstruct one from another. Let us take the Earth again, as it has been our example for all concepts in this chapter. We have already introduced the field line picture in section 9-1-3-3. We know that the equipotentials in this case (shown as dashed lines) are all spheres as U only depends on r for a spherical mass. But even if we did not know this, we would be able

to reconstruct the equipotentials by drawing lines perpendicular to the field lines as shown below:



While *every* sphere is an equipotential, we choose to only draw selected equipotentials. We draw equipotentials that are equally separated in *potential*, but not in space. For example, the equipotentials shown in the above figure may be -6×10^6 J/kg for the one closest to the Earth, -5×10^6 J/kg for the middle equipotential and -4×10^6 J/kg for the outermost equipotential. We see that even though that the equipotentials drawn change by 1×10^6 J/kg, they are not spaced evenly. As the equipotentials get further apart, we have to travel further with (or against) the field to get the same change in potential. This tells us that the field gets weaker. Finally, notice that if we only had the equipotentials (as on the diagram above on the right) we could reconstruct the field lines.

Another example of equipotentials is the example of a topographical map, as shown as an example of a "height field". The contours show locations of constant height, and close to the Earth's surface we have

$$U = \frac{\mathrm{PE}_{\mathrm{grav}}}{m} = gh$$

so lines of constant height h are also lines of constant U. The closer the equipotential lines are, the steeper the slope and the greater the combined force of the ground and gravity are on an object.

9-1-4-3 Force and equipotentials

So far we have the idea that the closer the equipotentials are, the stronger the field. In fact we can make this relationship precise:

$$|\boldsymbol{g}| = \left|\frac{\mathrm{d}U}{\mathrm{d}r}\right| \approx \left|\frac{\Delta U}{\Delta r}\right|.$$

Here Δr is the *shortest* distance from a point on one equipotential to a point on a neighbouring equipotential. The direction of the field is in the direction that the equipotentials are closest together.

In terms of "equipotentials" on the hill this means that the steeper the hill, the smaller Δr, and the larger acceleration an object placed on this hill would experience. This is not an explanation of why the acceleration of an object would be greater – for that you should go back to force diagrams of a ball on a hill – but it is a convenient way of mapping the acceleration that a ball would feel. This is analogous to writing the gravitational equipotentials – they are a convenient description, but they do not explain why the field is the way it is. While the hill serves as a good analogy, it is important to note that we are looking at the combined effect of gravity and the ground when discussing the acceleration of a ball. *The gravitational field does not change significantly on a hill!* An example with just the gravitational field is given in example # Section 9-1, ex. #2.

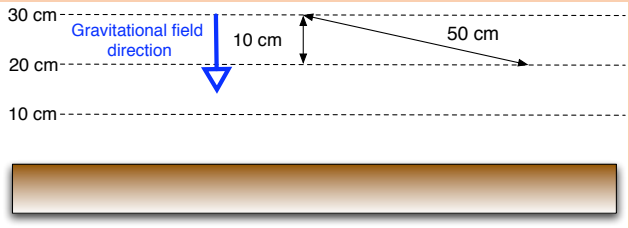
Example #2:

Two equipotentials close to the surface of the Earth have a potential difference of 1 J/kg. How far apart are they?

Solution:

We begin by taking $\boldsymbol{g}_{\text{Earth}} = 10\,\text{m/s}^2$. We are interested in making steps of $\Delta U = 1\,\text{J/kg}$ every equipotential. This tells us that the equipotentials are separated by

$$\Delta r = \frac{|\Delta U|}{|\boldsymbol{g}_{\text{Earth}}|} = \frac{1\,\text{J/kg}}{10\,\text{m/s}^2} = 0.1\,\text{m}$$



When we say the separation of the equipotentials is 10 cm, we mean that the closest the next field line gets to the point in which we are interested is 10 cm. The fact that the distance can get longer (for example, 50 cm as shown in the diagram) is completely irrelevant. The direction of the field is

perpendicular to the equipotentials, going from a high equipotential to a low equipotential. In this case, the equipotentials are closest vertically and the potential decreases in height, leading to the (already known) conclusion that gravity points down.

Example #3:

The 1 J/kg equipotentials at the surface of Pluto are separated by 1.6 metres. Is the gravitational field on the surface of Pluto stronger or weaker than the gravitational field at the surface of the Earth?

Solution:

The equipotentials are spaced further apart (larger Δr) for the same ΔU. Therefore the gravitational field at the surface of Pluto is *weaker* than the gravitational field at the surface of the Earth.

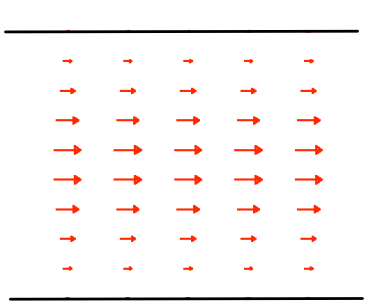
Notice that this does not explain *why* Pluto has a smaller gravitational field than Earth. To figure that out we would look at Pluto's mass and size compared to Earth. But if someone has already calculated the field or the equipotentials for us, we can still use that information to answer questions.

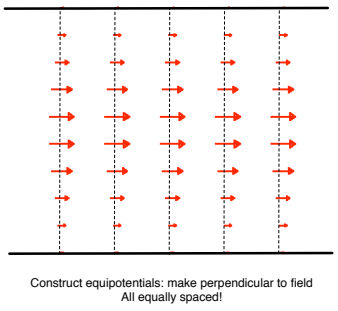
9-1-4-4 Potentials don't always exist*

As we learned in Physics 7A, the work done moving an object around is not a state function. This meant that the amount of work it took to move an object from one location to another could depend on more than the initial and final points; it could also depend on how you went from the initial to final point!

If the amount of *potential energy* a charge loses depends on the path taken, then it would seem that the change in *potential* would depend on the path taken as well. But this makes no sense: the potential is *defined* at a particular point without a reference to a path. To calculate the change in potential we simply take the difference of the potential at the two end points! Therefore, if the change in potential energy of an object depends on the path taken, then the potential does not exist! (There are other things that can happen that can prevent a potential from existing as well).

Let us show an example where it is impossible to construct a potential by showing it is impossible to construct an equipotential. Consider the vector map shown on the left:





This is not completely artificial; as we learned in 7B the water velocity in a real pipe is like this. The friction on the sides reduce the velocity to zero, and the velocity is highest in the centre. Our first attempt at constructing equipotentials will use the rule that the equipotentials are always perpendicular to the field lines. Because all the field lines are horizontal, our equipotentials will be vertical, as shown above on the right. But there is a problem with this; as the field gets weaker (toward the edges) the equipotentials should be getting further apart. But the equipotentials cannot stay perpendicular to the field *and* get further apart: the first condition requires them to be vertical, the second requires them to bend. In this simple example, there are no equipotentials! This tells us that the potential does not exist either, and with a little bit of work you could show that the work required to move a charge through an electric field like this would depend on the path taken.

The fact that we have two requirements that are not both automatically satisfied tell us that equipotentials only exist in very special circumstances. The cases that we are concerned with where potentials (and equipotentials) *don't* exist are

- A changing electric field (this is actually essential to induction, introduced in §9-3-4). This is because the electric field creates closed loops, as pointed out in §9-4
- The magnetic field *never* has a potential, as the magnetic field cannot do work. We will learn in §9-3-2-1 that the magnetic force is always *perpendicular* to the direction a charge travels, so $W = |\boldsymbol{F}_{\|}|\,\Delta x$ must be zero.

9-1-5 Superposition

So far the general principles of fields have been introduced by using a spherical or point mass as an example. For this case, we have a general formula for

the gravitational field

$$\boldsymbol{g}_{\text{mass M}} = \frac{GM}{r^2}$$

where the direction is always pointed toward the mass M.

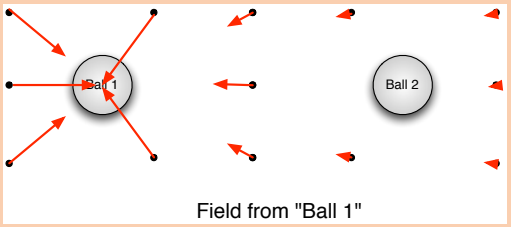
If we have two point-like or spherical masses, then we use *superposition*. When we used this word with waves in section 8-2, this meant that we added the two waves together to find the total displacement. When we use it for fields this means that we figure out what the field from the first mass is at the point of interest, then what the field from the second mass is *at the same point* and add them together. It is **critical** to remember in the last step that we are adding them as vectors[4]! Another important point to remember: We will *never* add field vectors at different locations. Any time we "superpose" anything, we will be looking at the effect from two (or more) *different sources* on the *same* location at the *same* time.

Example #4:

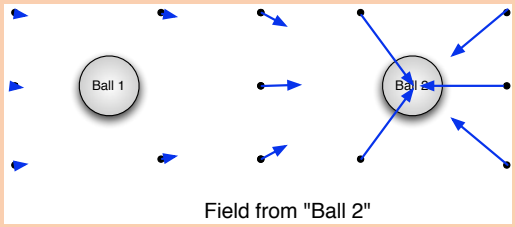
Draw a vector map for the gravitational field of two separated spherical balls of equal mass. i.e. pick a reasonable number of points at which to evaluate the field.

Solution:

Let us start by drawing two pictures of the situation. On the picture on the left we only include the field from "ball 1", and on the right we only include the field from "ball 2". These sketches are only rough, but they show that as we get further from the source the field gets weaker, and the direction of the field is always toward the source.



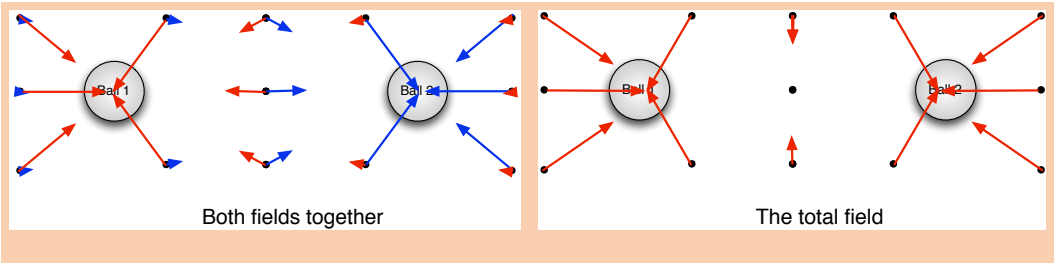
Field from "Ball 1"



Field from "Ball 2"

Now we have to put these two fields together, and add them the fields as vectors. On the left hand side are both of the fields before doing the vector sum, and on the left is the total vector field (i.e. after doing the vector sum).

[4]Technical aside: because superposition is the most reasonable thing to do, it is sometimes easy to forget that we are assuming something here. Superposition of (weak) gravitational fields is an *experimental* conclusion, not one that can be made by purely logical thought.



We now have enough information to deal with an arbitrary mass distribution. We can think of taking any distribution of mass and breaking it into a lot of different point masses. We can figure out the field from these point masses on any point, and then add these contributions together to find the total gravitational field. This would involve a lot of work, but at least we know how to do it in principle.

9-1-6 Relationship between concepts

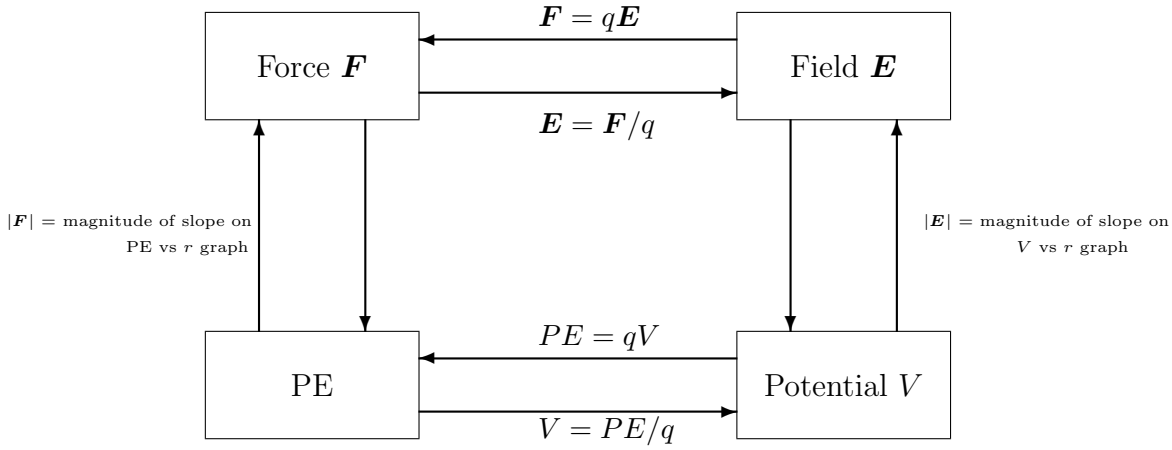
While the gravitational and electric fields are very different, they share many of the same relationships. Here some diagrams summarise how to translate between our new concepts of fields and potentials back to the familiar concepts of forces between two objects and the potential energy between those objects. These diagrams are different but have very similar structure. The main difference (other than the names of the field) is that in the electric case to move "horizontally" between old and new concepts we use the electric charge, while for the gravitational field we use the mass.

On the other hand, the analogous diagram for the magnetic field is very different. That is because there is no analogue of magnetic potential energy.[5]

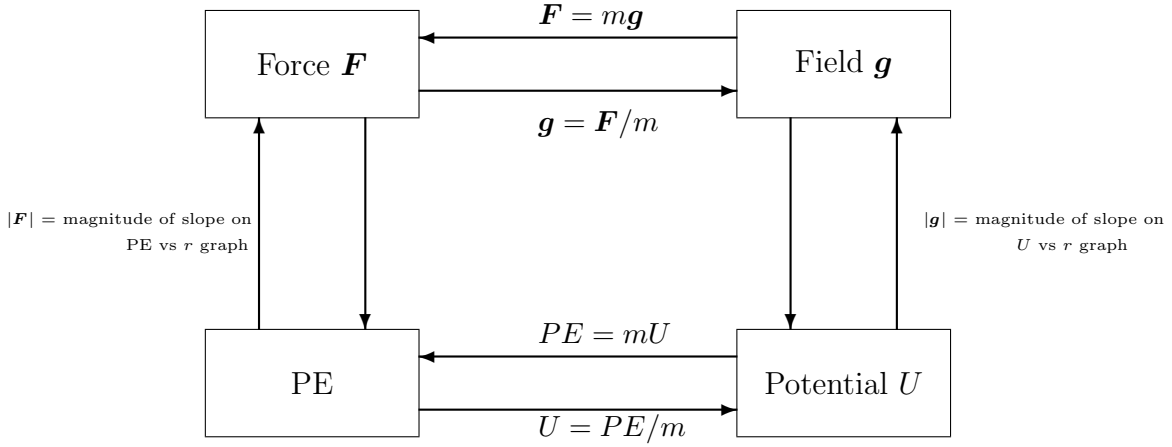
Electric field/force/potential relationships

Potential exists for time-independent electric fields only

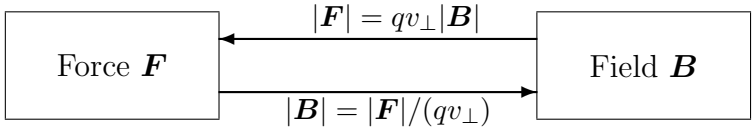
[5]Technical aside: The magnetic field can store energy, as can the electric field. This is not the same as the potential energy a charge stores in the field, as this represents the ability of the field to do work on the charge.



Gravitational field/force/potential relationships



Magnetic field/force relationships



The magnetic field has no potential, and also depends on the velocity. The direction of the magnetic force is **not** is not in the same direction of the magnetic field; the relationships above are for the *magnitudes* only. Finding the direction of the magnetic field is a little bit involved, and this diagram should make more sense after studying the magnetic field in chapter 9-3.

9-1-6-1 Relationship between representations

We now know three different representations of a field: using a field map, field lines or equipotentials.

- **Field map**
  Introduced in section 9-1-3-3
  In the field map representation, we calculate the vectors at certain points at a given time. Usually these points are taken to lie on a grid. We then draw vectors *to scale* to indicate the direction and magnitude of the field. This is the most direct representation of the field, but when the field varies wildly in magnitude it is difficult to make the small vectors large enough to be seen while keeping the large vectors from extending too far. It also involves a lot of work to calculate.

- **Field lines**
  Introduced in section 9-1-3-3
  This representation tries to fix some of the problems with the vector map representation. Here the field lines are joined together to make continuous field lines. To recover the direction of the field at a particular point, we need to take a tangent of the field at that point. The spacing between the field lines indicates the magnitude of the field; the denser the lines the greater the magnitude of the field.

- **Equipotentials**
  Introduced in section 9-1-4-2
  The lines of equal potential. These are always at 90° to the field lines, and we don't lose potential energy moving along them. (Note: these do not exist for magnetic fields). Regions where the field is strong the field lines are close together.

The table below summarises how to read each one, and how to go from one representation to another.

Quantity	**How to read vector field (e.g. E, g) with**		
	Field map	Field lines	Equipotentials
Magnitude	Length of vector starting at point	How dense field lines are.	How far apart two neighbouring equipotentials are
large field	Long arrow at point	dense field lines	close equipotentials
small field	Short arrow at point	sparse field lines	separated equipotentials
Direction	Direction of arrow at point	Direction of tangent to field line at point	Direction of normal to field line at point
→ field lines	Connect arrows by curves	—	Take equally spaced normals along a equipotential. Then extend these normals, bending them so they are always normal to any equipotential you cross.
→ equipotentials	Make into a field diagram first (as above)	Take a normals along a field line. Extend these normals, bending them so that they are always perpendicular when crossing another field line.	—

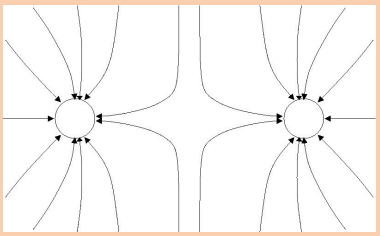
Example #5:

Show the gravitational field lines and the equipotentials for two separate balls of equal mass. You can start from the field map given in Section 9-1, ex. #4.

Solution:

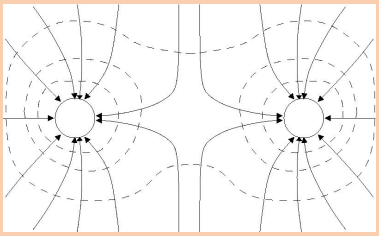
Field lines:

To go from the vector map to the field line map, we try and connect the field line arrows in a smooth manner. A rough drawing is shown below:

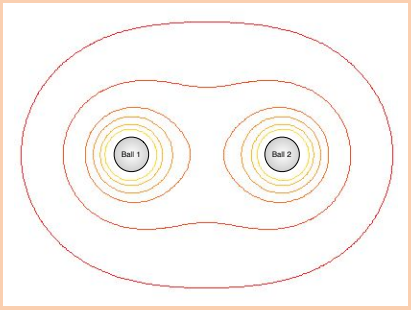


Equipotentials:

To find the equipotentials we start on a field line and draw across it perpendicular. Then we continue the field line, and keep bending it to ensure that it is perpendicular to the next field line it crosses. A *rough* sketch to do this is shown, where the dashed lines are the equipotentials.



How close is this to the "true" answer? A computer generated answer shows us the true equipotentials:



We see that our answer captured most of the correct answer. One of the reasons potentials are so useful for physicists is that they are easy to calculate exactly and generate plots like the one above. In this course, which does not emphasise computation, there is no real advantage to constructing the equipotentials first and then the field lines or vice-versa.

9-1-6-2 Relationship between fields and waves

The examples of the gravitational and electric fields shown in this chapter have all involved stationary sources, and hence the fields have not changed in time. However if we get the sources to move, then the field at a particular location starts to change in time as well. If we disturb the charges, this disturbance propagates outward like a wave pulse. If we make the field vary *periodically* then we get waves in the field that are identical to the waves we studied in the beginning of the course. In the case of the electric and magnetic fields this sort of oscillation is what we commonly call "light". This fact is so important that it bears repeating:

- A field can oscillate, producing a travelling wave.

We will return to this idea in more detail after we have introduced electromagnetic waves in §9-4-5.

Up until this point we have been using a field as a convenience, but it seemed like we could always just use the direct model of forces and not have to deal with it. The fact that we can get fields to generate waves that transport energy (and momentum) tells us that there is more to it than that. If we did not count the field as real, one system would lose energy, that energy would remain lost, and sometime later another system may gain energy – throwing energy conservation out the window! If we consider the field as a physical entity in its own right, rather than just a trick, the description changes: now a system can *transfer* energy to the field, the field transports that energy as a wave, and then the field *transfers* the energy to another physical system. We see that fields are more than a trick, but are required if we want to preserve the conservation of energy!

9-1-7 Summary

This chapter has introduced many new ideas related to fields, using the gravitational field as the primary example. In the upcoming chapters the electric and magnetic fields are also discussed, so do not be concerned that their treatment in this chapter was brief. The main concepts introduced were:

1. That a field was a physical quantity that was not in a specific place, but spread out. A field can also change its value in time. It is different from a wave because a wave *must* change its value in time.

2. A force is an interaction between *two* objects. A field is created by a *single* object.

3. The fields we will be interested in are *vector fields*, meaning that the field has a direction and magnitude at every location in space.

4. When using the field model, one object (the "test" object) feels the field created by everything else (the "source" objects). To find the field created by everything else we use superposition.

5. An object does not feel its own field.

6. Representations of vector fields:

    - A vector map:
      A "snapshot" of the field vectors at a particular time. (e.g. wind map)

- Field line map:
  A "snapshot" of the field, but with continuous lines. The direction of the field at any point is tangent to the field lines. The strength of the field lines is determined by how close together the field lines are. If they are bunched up the field is strong, if they are spread thinly the field is weak.
- Equipotentials: (not for magnetic fields)
  If a potential exists, then the equipotentials are the directions where we would not have to do any work to move the object. That is, we are not going "with" or "against" the field. The equipotentials are always at 90° to the field lines. (e.g. contour map as equipotentials of gravitational field)

7. An electric field line starts on a positive charge and end on negative charge.

8. A gravitational field line start at infinity, and end on a mass (no "negative masses").

9. Magnetic field lines close in on themselves; they never start or end.

9-1-8 Exercises

1. A 1 kg mass and a 5 kg mass sit in the *same* gravitational field at the *same* location. For which mass is $\boldsymbol{F}_{\text{grav field on mass}}$ greater?

2. Estimate the force of the Moon on you as you stand on the surface of the Earth. Also estimate the force of the sun on you as you stand on the surface of the Earth? Will the answer change between day and night, and is this change significant? Which is stronger – the pull of the Moon or the pull of the Sun?

3. Where is the location which has equal magnitude for $\boldsymbol{g}_{\text{Earth}}$ and $\boldsymbol{g}_{\text{Moon}}$? Draw a picture of the Earth and the Moon and indicate the approximate location with an "x". Check your answer: should the "x" be closer to the center of the Earth, or closer to the center of the Moon?

4. If we have a planet half the radius of the Earth, but the same mass will the gravitational field on the surface of this planet be greater than, less than or equal to the gravitational field on the surface of the Earth?

5. If a ball of mass 2 kg is thrown vertically upward from the surface of the Earth at a speed of 10 m/s how high would it go before stopping? Ignore air resistance. Hint: Use an energy conservation argument. Should you use $\mathrm{PE}_{\mathrm{grav}} = mgh$ or $\mathrm{PE}_{\mathrm{grav}} = -GmM/r$? Do you get a significantly different answer by making a different choice?

6. If a ball of mass 2 kg is thrown vertically upward from the surface of the Earth at a speed of 10,000 m/s (i.e. very fast) how high would it go before stopping? Ignore air resistance. Hint: Use an energy conservation argument. Should you use $\mathrm{PE}_{\mathrm{grav}} = mgh$ or $\mathrm{PE}_{\mathrm{grav}} = -GmM/r$? Do you get a significantly different answer by making a different choice?

Unit 9:

9-2: Electric fields

Almost all of the main ideas about electric fields were presented in the previous section about generalised fields (section 9-1). If you have not read that section, return to it and read it now.

This section will help you apply the ideas that were somewhat familiar in a gravitational context to a new arena. By the end of this section, you will understand the ideas of electric charge, electric force, electric field, electric potential energy, and electric potential.

9-2-1 Electric charge

Previously, we said that the Earth's gravity is stronger than the Moon's gravity. We were able to quantify this statement by determining that, at its surface, Earth has a greater gravitational field than the Moon. We were able to quantify this statement by determining that, at its surface, the Earth has a greater gravitational field than the Moon. This meant that a given mass felt a greater gravitational force on the surface of the Earth than on the surface of the Moon. We learned in section 9-1 that the strength of the gravitational field was determined by a spherical mass and the radius of the body, and numerically it can be shown that the Earth's gravitational field is greater.

To make an analogous argument about electric fields, we must determine the electric analogue to mass. It will probably come as no surprise to you that this analogue is charge. The greater an object's charge, the greater electric field it will produce at the same distance from the object. Nearly all our ideas about mass can transfer to our discussion of charge, with one clear difference. Unlike mass, charge comes in two types, positive and negative.

Electric charges are measured in Coulombs, abbreviated C. The electron has a charge of -1.602×10^{-19} C (the proton's charge has the same magnitude, positive sign). All atoms, molecules, or macroscopic charged objects get their charge from either an excess or deficit of electrons compared to the number of protons they have; a charge of 1.602×10^{-19} C is thus considered fundamental, as all observed particles or objects have some integer multiple of this value.

9-2-2 The electric force

We can determine the electrical force between two charges in one of two ways: the direct model or the field model. In the direct model, we determine the magnitude of the electric force without any reference to the field:

$$\text{Charge \#1} \overset{\text{creates force on}}{\Longrightarrow} \text{Charge \#2}$$

From Newton's third law, we know that charge #2 simultaneously exerts a force of the same magnitude on charge #1. In the direct model the two charges are treated the same way.

The magnitude and direction of the electric force from a *point* source charge Q on test charge q, with a distance r between their centres, is given as:

$$\boldsymbol{F}_{\text{electric } Q \text{ on } q} = \begin{cases} \text{Magnitude} = kQq/r^2 \\ \text{Direction} = \text{attractive if q and Q have the opposite signs} \\ \qquad\qquad \& \text{ repulsive if q and Q have the same sign} \end{cases}$$

The constant $k = 9\times10^9$ Nm^2/C^2 converts everything to the proper units of Newtons. Compare the direct model for electrical forces to the direct model for gravitational forces. Note that in both cases, the strength of the force depends on the inverse square of the distance between the objects' centres. That is, if we double the distance between two charged objects, both the gravitational force and the electrical force will be one fourth the previous values. Also note that mass and charge enter the equation in an identical fashion.

If you are asked to draw a force diagram for an electrical setup, note that you determine the direction(s) and the magnitude(s) of all relevant forces in separate steps. To determine the direction of the electric force, we do not use equations. Instead, recall from Physics 7A that like signs repel but opposite signs attract.

9-2-3 The electric field

Another way to determine the electric force is via the field model. Throughout this quarter, we will use the field model more frequently than the direct model. In the field model, we analyse the interaction between two charges in two steps:

$$\text{Charge \#1 (source charge)} \overset{\text{creates field}}{\Longrightarrow} \boldsymbol{E} \overset{\text{exerts force on}}{\Longrightarrow} \text{Charge \#2 (test charge)}$$

Instead of directly calculating the force the source charge #1 exerts on the test charge #2, we think of the source charge as creating an electric field $\boldsymbol{E}$, and then this field exerting a force on the test charge. Once we have determined the electric field produced by the source charge, its job is done, and we can determine the force on test charge entirely from the field $\boldsymbol{E}$. As with the gravitational field $\boldsymbol{g}$, the electric field $\boldsymbol{E}$ exists in all points of space, and may or may not change over time. In addition to having a value at all points of space, the field also has a direction. The property of having both a magnitude and direction at every point makes this field a vector field. We can represent these properties with either field vectors or field lines (more to follow).

The first step in the field model is to determine the field created by the source charge, and this step does not involve the test charge in any capacity. Likewise, the second step in the field model is to determine the force created by the field. At this stage, we no longer consider the source charge but instead consider the field the source charge produced . In the direct model, we consider the interaction between the two charges directly. In the field model, we consider each charge in a separate step.

Now that we have defined the ideas of electric fields and forces, let's explore the relationship between the two more exactly. We previously determined that the electric force can be calculated from:

$$|\boldsymbol{F}_{\text{electric } Q \text{ on } q}| = \frac{kQq}{r^2}$$

In defining our electric field, we must ensure that the field does not depend on the test charge. Consider Q to be the source charge and q to be the test charge. The simplest construction we could make without the test charge is kQ/r^2, which would be the electric force on a test charge of one Coulomb. To determine the force on any other test charge, we would take the value kQ/r^2 and multiply by the magnitude of the test charge. This idea should

sound very familiar, as it is exactly what we did in defining the gravitational field!

Thus, we find that the electric field is given by:

$$\boldsymbol{E}_{\text{of source charge } Q} = \begin{cases} \text{Magnitude} = |kQ/r^2| \\ \text{Direction: in towards } -Q\text{; out away from } +Q \end{cases} \tag{9-2.1}$$

The units of the electric field are Newtons/Coulomb. There are other equivalent units such as Volts/meter that we will see later (§9-2-5-1). The magnitude of the $\boldsymbol{E}$ field is an absolute value–its just a length of a vector. The electric field's direction at a point is in the direction a *positive* charge would feel a force. A negative charge feels a force in the *opposite* direction to the electric field. As a consequence of this convention, the fact that a positive charge repels other positive (test) charges means that the electric field *starts* on a positive charge and points away. A negative charge would attract a positive (test) charge, and therefore negative charges create electric fields that point inward. The total electric field is taken by combining the electric fields of all the source particles and superposing them (see §9-1-5).

To determine the strength of the electric force, multiply by the magnitude of the test charge:

$$\boldsymbol{F}_{\text{field on } q} = q\boldsymbol{E}$$

The units work out nicely to give Newtons, the expected units of force. If we combine the two steps of the field model of electric forces, we find:

$$\left|\boldsymbol{F}_{\text{field from } Q \text{ on } q}\right| = |q\boldsymbol{E}_{\text{from } Q}| = \left|q\left(\frac{kQ}{r^2}\right)\right| = \left|\frac{kQq}{r^2}\right|$$

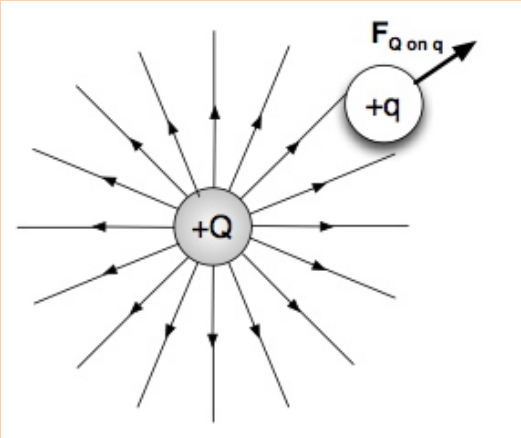
This is merely the direct model of electric forces. It is a nice result that we mathematically obtain the same answer for the magnitude of the force on q using either method (direct or field).

Example #1:

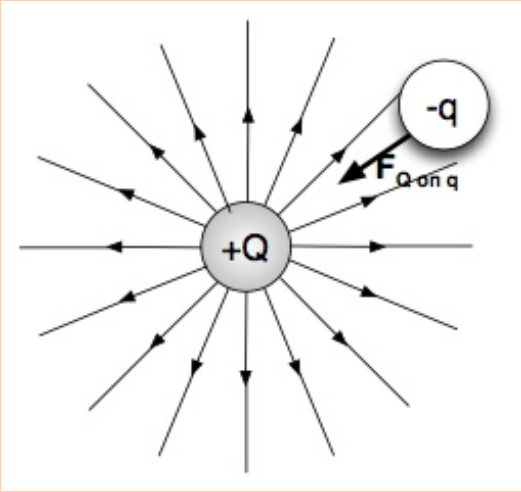
In the direct model of force, we found that the force is attractive when the charges are different signs and repulsive if the charges are alike. How can we use the electric field to determine the direction of the force?

Solution:

We know that positive charges create electric fields that point outwards in all directions of space. If we place a second positive charge in the field, we will have a repulsive force that also points away from the source charge:



If instead we place a negative charge in the field, we find an attractive force that points toward the source charge, in the opposite direction to the field:



At least in the case of two charges in space, we find that that the force is in the same direction as the field for positive source charges, and opposite the field for negative charges.

It turns out that the relationship between force and field direction explored in example 5 generalises. No matter how complicated the charge

configuration, if we know the direction of the electric field, we can easily determine the direction of the electric force.

$$\boldsymbol{F}_{\text{of field on charge } q} : \begin{cases} \text{Magnitude} = q|\boldsymbol{E}| \\ \text{Direction} = \text{along } \boldsymbol{E} \text{ field vector for} + q; \\ \qquad\qquad\quad \text{opposite } \boldsymbol{E} \text{ field vector for} - q \end{cases}$$

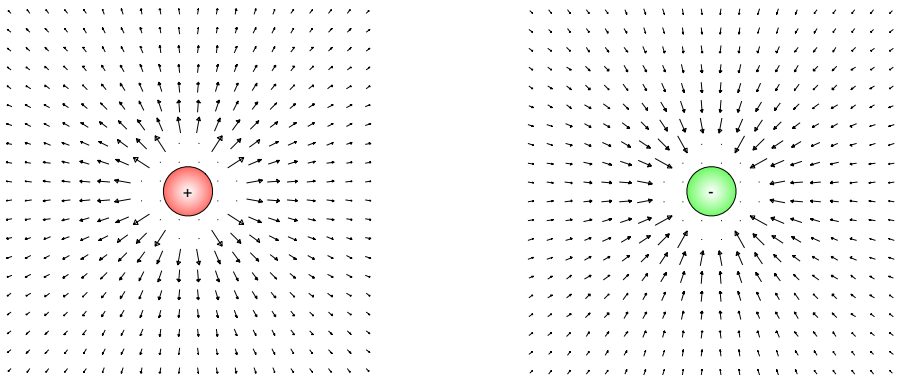
Before proceeding, we should pause for a moment. We have provided an equation to calculate the electric field created by a single charge. If we would like to find the electric field created by two charges at a certain location in space, we know we can use the principle of superposition (as in §9-1-5) to add the various fields at that specific location. Note that we must add the fields as vectors; simply adding the fields' magnitudes is not correct (depending on orientation, adding two fields with equal magnitude can result in a net field of zero, of double the individual fields, or anything in between). We could extend this process to calculate the field from three charges, 10 charges, or 10,000 charges. Every unique configuration of charge will have an equation for the total electric field. If the objects are *not* spherical or points, the electric field is *not* given by equation (9-2.1)

9-2-3-1 Representing the electric field

Throughout §9-1 on fields, we explored three ways to represent fields. In this section, we will expand on two of these representations for the case of electric fields: vector map and field line map.

Vector map

In order to make a full vector "map" of the electric field surrounding a source charge Q, we must evaluate the magnitude and direction of the electric field vector at each and every point in space. As this is impractical, we settle for calculating the field at various evenly spaced points on a grid surrounding the source charge, but even this can be a daunting task! You should begin to be comfortable with calling this "map" of electric field vectors the "$\boldsymbol{E}$ field." Some very long $\boldsymbol{E}$ vectors (i.e., at points very close to the charge Q) have been omitted for clarity. One advantage of the vector map representation is that it highlights specific field vectors at specific points in space. Compare the below maps to those for gravitational fields in section 9-1-3-3.



Field lines

A second way we can represent the electric field is through field lines. Field lines are a graphical "shortcut" to represent electric fields, and are drawn by connecting field vectors together. Electric field lines always start from a positive charge and end on a negative charge (or extending indefinitely into space). The direction of the field vector can be determined from the field line; the field vector is tangent to the field line (see diagram). (diagram showing a complicated field line map, showing several sample field vectors)

Example #2:

Explain how you can determine the strength of the electric field using a) the vector map representation and b) the field lines representation. You may find the relationships in section 9-1-6-1 useful.

Solution:

a) In the vector map representation, a field vector is provided at many sample locations throughout a region. Locate the field vector closest to the point of interest. The length of the vectors indicates magnitude of the electric field at that location. Note that the actual scale of the vectors is somewhat arbitrary and chosen for clarity (that is to say, is 1 cm equal to 10 Newtons/Coulomb? Or 1,000 N/C?) The relative length of different arrows determines the relative strength of the electric field. The actual magnitude could be determined from the scale.

b) In the field lines representation, the density of the field lines determines the strength of the electric field. Locate the point of interest, and see how many field lines are nearby. Consider placing a quarter on the paper at that point. How many lines does your quarter cover? How does this compare to how many lines your quarter would cover elsewhere on the page? The relative density of the field lines determines how strong the field is, though

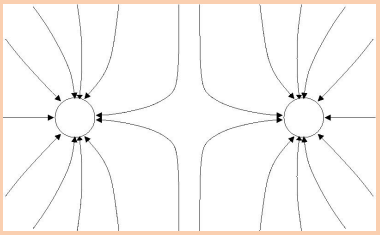
it not possible to determine an actual magnitude direction from the field lines representation.

Example #3:

In Section 9-1, ex. #4 and Section 9-1, ex. #5 from the section about fields, you developed a vector field map and a field line map for gravitational field created by two separated spherical balls of equal mass. Now suppose that each ball carries a net charge, and that furthermore the electric field they generate can be represented by the exact same maps. Determine the signs of the charges.

Solution:

Looking at the field line map (reproduced below), it is clear that the field lines end on the charges. Field lines end on negative charges, so each charged ball must carry a negative charge.



9-2-4 Electric potential energy

Believe it or not, you already know a great deal about electric potential energy, which you studied extensively in Physics 7A. For instance, it takes energy to move two like charges closer together. We can model the process of moving charges closer together with the following energy interaction diagram:

$$W = \Delta \mathrm{PE}_{\mathrm{electric}}$$

Now let's imagine starting with a positive charge and a negative charge very far apart, and allowing them to come nearer. The charges are attracted to one another. As they come nearer, they speed up due to this attraction. We can represent this interaction with the following energy system diagram.

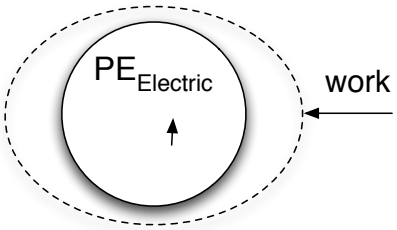


Figure 9-2.1: It takes work to more two like charges closer

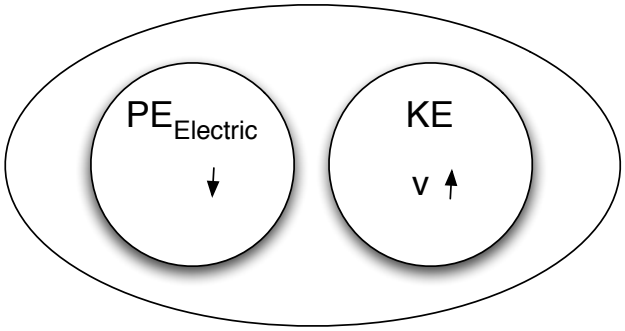


Figure 9-2.2: As unlike charges near, $\text{PE}_{\text{electric}}$ is transferred into KE.

$$\Delta E_{\text{tot}} = 0 = \Delta\text{PE}_{\text{electric}} + \Delta\text{KE}$$

Like force, potential energy is an interaction and requires at least two charges. It makes no more sense to talk about the potential energy of a 45 Coulomb ball than it does to talk about the force of a 45 Coulomb ball. To have either a force or potential energy, a minimum of two charges are required.

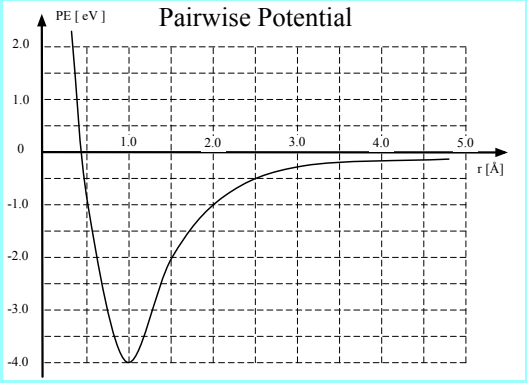
As both force and potential energy are interactions and require at least two charges, one might expect them to be related in some way. Indeed, they are. Their relationship was studied in Physics 7A: how quickly the potential energy changes as position changes determines the magnitude of the force.

$$|\boldsymbol{F}_{\text{something on object}}| = \left|\frac{\text{d PE}}{\text{d}r}\right|$$

Recall that graphically, evaluating the derivative at a certain location is equivalent to finding the slope of a PE vs r graph at that location. If the potential energy is changing rapidly, then the graph will be steep, the slope big, and the force at that spot great. If these ideas are unfamiliar to you, consult appendix D of this volume or your introductory calculus text.

Example #4:

As studied in Physics 7A, the attraction between two atoms can be modelled as a Lennard-Jones interaction. Determine the force at a separation distance of a) 1.5 Angstroms and b) 4 Angstroms.

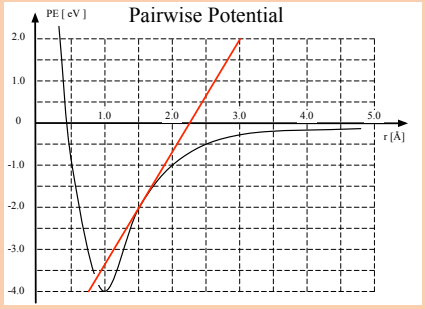


Solution:

We are asked to evaluate the force at two different locations. To do this, we must draw a tangent line at each location (1.5 Å and 4 Å) and calculate each line's slope. This will determine the magnitude of the force. We must also determine the direction of the force.

a)

A tangent line must have the same slope as the original function. Taking a ruler and matching the slope, we find



The task now before us is to calculate the slope of this line. Between the 0.75Å and 3Å locations, the potential energy changes by 6 eV. The slope is rise over run, or

$$\text{slope} \approx \frac{6\,\text{eV}}{(3\text{Å} - 0.75\text{Å})} = 2.67\,\text{eV/Å}.$$

While we have certainly determined the magnitude of the force, the units of force we are accustomed to are Newtons, not eV/A. Before we call our work complete, we should convert to Newtons.

$$1\,\text{eV} = 1.6 \times 10^{-19}\text{Joules, and } 1\,\text{Å} = 10^{-10}\,\text{meters}$$

$$\left(2.67\frac{\text{eV}}{\text{Å}}\right)\left(1.60 \times 10^{-19}\frac{\text{J}}{\text{eV}}\right)\left(\frac{1\,\text{Å}}{10^{-10}\,\text{m}}\right) = 4.3 \times 10^{-9}\,\text{J/m} = 4.3 \times 10^{-9}\,\text{N}.$$

As far as direction goes, at the 1.5 Å mark, the atoms are attracted. The force is to the left. The force will always act to decrease the potential energy.
b)
At a separation of 4 Å, the potential energy graph is nearly flat. Thus, the slope approaches zero, and so does the force.

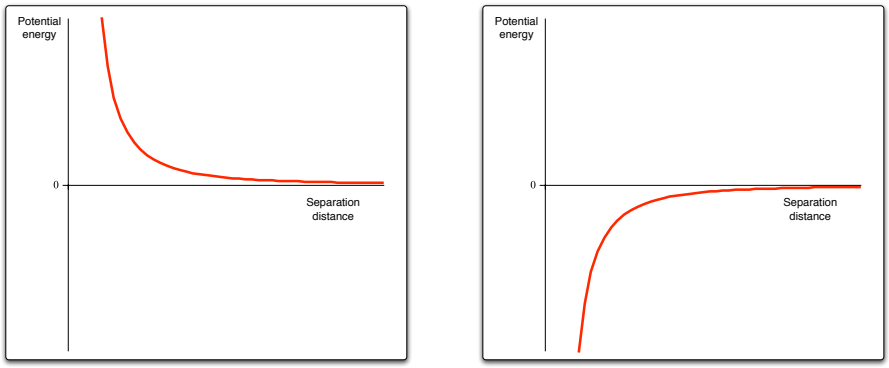
There is no single equation for potential energy. For instance, the potential energy of two atoms interacting (as in the Lennard-Jones interaction, above) is different than the equation for two single charges interacting. We have discussed the electric field created by a single charge, and the electric force between two charges. We now use this prior work, along with the relationship between force and potential energy, to determine the potential energy of two charges interacting.

The magnitude of the force between two spherical charges is $|kqQ/r^2|$. We know that the force is equal to the derivative of the potential energy with respect to position; $|\boldsymbol{F}_{Q \text{ on } q}| = |\text{d PE}/\text{d}r|$. We would like to know the potential energy as a function of position. Our question becomes, "what function has a derivative of $-kqQ/r^2$?" An integral later, we find that the potential energy of charges interacting is given by:

$$\text{PE} = \frac{kqQ}{r} + \text{constant}$$

Our only prior requirement was that the derivative (or slope) of the potential energy with respect to position gives us force. The constant does not change the derivative (slope). It is up to us to determine the value of the constant, and in so doing, the zero-point for potential energy. This idea should be familiar from Physics 7A. It is convenient to choose charges separated by a long distance to have zero potential energy. With our current equation for potential energy, PE $=kqQ/r+$ constant, when we consider very large separation distances, we have PE= constant. To make this correspond to having no potential energy, the constant must be equal to zero.

With two signs, there are three different combinations of charges: both positive, both negative, one charge of each sign. As far as the potential energy is concerned, either case of like charges results in a the same potential energy for all separation distances, so only two cases need be treated: like charges or different charges, as in figure 9-2.3.



(a) Like Charges (b) Unlike Charges

Figure 9-2.3: Potential energy graphs for spherical charges. a) Like charges have positive potential energy. b) Unlike charges have negative potential energy.

For like charges, the potential energy is always positive. Considering the total mechanical energy, (PE + KE), and knowing that kinetic energy is always positive in classical systems, the total mechanical energy must be positive as well. If the charges are initially moving toward one another, energy transfers from KE to PE until finally all of the energy is in PE, at which time the particles briefly stop, turn around, and move apart.

For unlike charges, there are two interesting cases. If the total mechanical energy is greater than 0, then the particles will have both kinetic energy and potential energy for all separation distances. If the total mechanical energy is less than 0, then the particles are confined to one another in a bound state. At these energies, the particles lack sufficient energy to escape their electric attraction. A common example of this phenomenon is the hydrogen atom, which will be analysed in more depth in the final chapter.

Test yourself:
Note for either potential energy graph, the PE gets flat for large separation distances and steep for small separation distances. What can you conclude from this information?

At this point, we have explored the ideas of charge, force, electric field, and electric potential energy. The last concept to tackle is the electric potential, corresponding to the gravitational potential U in section 9-1-4-1.

9-2-5 Electric potential

Electric potential is different from electric potential energy!
Although force depends on having at least two charges, we found it useful to define a field $\boldsymbol{E}$ so that we could easily determine the force on any given point charge. Like force, potential energy depends on at least two charges. We can define a new quantity, electric potential, to tell us something about energy that can be determined by source charge(s) alone.

The electric potential tells us how much potential energy a point particle would have at a certain location. Specifically, the electric potential tells us how much potential energy a one Coulomb charge would have. Like energy, the electric potential is a scalar, having magnitude but no direction. Because the electric potential is defined at all points in space, it meets the qualifications of a field, and we can describe the electric potential with a scalar field.

We will define the electric potential in an analogous way to how we defined grativational potential in 9-1-4-1.

$$\mathrm{PE}_{\mathrm{electric}} = q_1 V_2$$

In this notation, V_2 is the electric potential created by charge two.

Knowing that electric potential energy is measured in Joules, and charge in units of Coulombs, it follows that electric potential V must have units of Joules/Coulomb. Electric potential is thus an energy density. We already have a quantity labelled V with units of J/C from Physics 7B: Voltage (recall that volts = J/C). As it turns out, voltage and electric potential are the same thing, and the terms will be used interchangeably from here out.

Test yourself:
In section 9-2-4 we discussed the electrical potential energy for interacting charges. In this section, we developed a relationship between $\mathrm{PE}_{\mathrm{electric}}$ and the electric potential for a point charge. Determine the equation for the electric potential from a point charge.

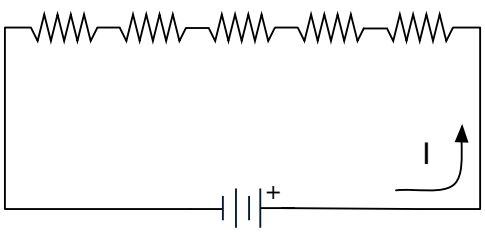


Figure 9-2.4: A simple circuit consisting of a battery and string of resistors.

Our prior experiences with electric potential were primarily in circuits. In a circuit, an electron gains a certain amount of energy in travelling across a battery from low voltage to high voltage. To draw an analogy with gravitational interactions, the process is similar to being lifted vertically and gaining gravitational potential energy as it travels from low to high gravitational potential.

In an electrical circuit, the voltage drops by IR across every resistor. Imagine we have a string of resistors hooked up to a battery as in Figure 9-2.4. In this case, the battery, wires, and resistor string together constitute the "source charge configuration." Before we mention a specific electron, we can discuss the electric potential at various locations in the circuit. However, we cannot talk about changes in electric potential energy until we introduce a specific electron, a test charge, and its path through the circuit. Electric fields and electric potentials depend only on the source; electric potential energy and force depend on the source and test charge.

Equipotential representation of V

We will continue with the example of a string of resistors hooked up to a battery. If all of the resistors have the same resistance R, then the voltage drop across each sequential resistor would be equal. To represent the electric potential, or voltage, we typically draw "equipotential maps" by connecting locations with the same potential. Furthermore, we typically indicate only equally spaced potentials (for instance, we *could* choose 3V, 2V, 1V, 0V, -1V, but we would *not* choose 3V, 2.5V, 1V, -3V). Figure 9-2.5 indicates places on the string of resistors with the same potential. Because all resistors have the same value, the voltage drop across each one is equal and we have met the requirements of having equipotentials spaced at regular intervals.

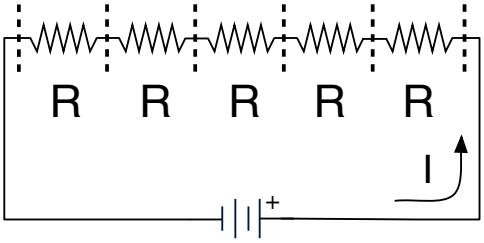


Figure 9-2.5: A simple circuit consisting of a battery and string of resistors.

For the circuit diagram 9-2.5, the equipotential representation adds little or no value to our prior circuit diagram representation indicating $\Delta V_{\text{a to b}}$, for instance. In circuits, we are only interested in what occurs along one dimension, the dimension of the wire. The real value of equipotentials occurs when applied to more dimensions, because then we can see the value at all points in space. By knowing the value of the electric potential, we know how much potential energy a charge would have it it were placed at that value. As we will see in a moment, information gleaned from the equipotential map also indicates the relative strength of the magnetic field, which can be used to determine the relative magnitude (and direction!) of the force on a point charge at a specific location.

As with the electric field, any configuration of charges has a unique electric potential. That is to say, the potential from a spherical charge, a capacitor, and a charged wire all have unique potentials and fields. You will have the opportunity to explore several of these during DL.

Example #5:

a) Draw three or four equipotentials for a proton, spanning the range between 10V and 30V. Be sure to include an appropriate scale on your equipotential map.
b) Using your equipotential map, estimate the value of the potential at the Bohr radius, 0.529 Å, which corresponds to the electron's average separation distance from the proton in the lowest energy state in the Hydrogen atom.

Solution:

a) The electric potential from a point charge Q is $\frac{kQ}{r}$. We are asked to span the range of voltages between 10V and 30V with three or four potentials. We know that our equipotentials should be chosen with equal voltage

differences. 10V, 20V, and 30V are our best bet.
Before blindly plugging numbers into the formula, let's think about what we expect to draw.

- The potential is proportional to $\frac{1}{r}$, so as we move further away from the proton, the potential will decrease. Applying this idea to our problem, we expect the 30V equipotential to be closest to the proton and the 10V furthest away.
- Furthermore, the potential decrease faster for smaller distances, such as from 1m to 2m, than it does for larger distances, such as from 99m to 100m (not convinced? Think of the graph of the function 1/r or plug the numbers into your calculator). Because of this, we expect the 20V equipotential to be closer to the the 30V potential than the 10V potential.

Now that we've thought through our expectations, we can look up the values of k ($9 \times 10^9 \frac{\text{Nm}^2}{\text{C}^2}$) and the charge of a proton (1.6×10^{-19} C) in the appendix. We calculate the value of r for each of the three voltages, starting with 10V:

$$\begin{aligned} 10\,\text{V} &= \frac{\left(9 \times 10^9 \frac{\text{Nm}^2}{\text{C}^2}\right)(1.6 \times 10^{-19}\,\text{C})}{r_{10\text{V}}} \\ r_{10\text{V}} &= 1.44 \times 10^{-10}\,\text{m} \\ 1.44\,\text{m} &= 1.44 \times 10^{-10}\,\text{m} \times \left(\frac{1\,\text{Å}}{10^{-10}\,\text{m}}\right) \\ r_{10\text{V}} &= 1.44\,\text{Å} \end{aligned}$$

Similarly for the other voltages,

$$\begin{aligned} r_{20\text{V}} &= 0.72\,\text{Å} \\ r_{30\text{V}} &= 0.48\,\text{Å} \end{aligned}$$

Next, we plot the values, being sure to indicate the scale. See Figure 9-2.6. Comparing our graph to our expectations, we find they match, so we can proceed to the second part of the problem.
b) Using our scale, we determine where the Bohr radius fits into the picture. The location is marked with a solid dot. Apparently, the value of the potential is between 20V and 30V, but much closer to 30V. We might estimate a value of 28V.
Though it is not asked for in the problem, we can calculate the value of the Bohr potential at that location. Plugging numbers into the formula, we find just over 27V. Our prediction wasn't bad!

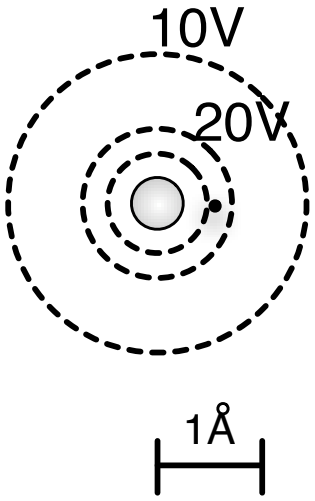


Figure 9-2.6: Three equipotentials created by a proton. The Bohr radius is marked with a solid dot

9-2-5-1 Relationship between *E* and V

Thus far we have focused our attention on the relationship between the potential energy and the electric potential V. We noted earlier that both $\boldsymbol{E}$ and V depend entirely on the source charges. We now explore the relationship between these quantities.

Let's think about two interacting spherical charges one last time. We can find the electric potential created by charge Q ($\frac{kQ}{r}$). Go from the electric potential created by Q to the electric potential energy of charge q, simply multiply by q, as in this section ($\frac{kqQ}{r}$). Using this, we take the derivative with respect to separation distance to find the force between the charges, as in section 9-2-4. ($\frac{kqQ}{r^2}$). From the force of charge Q on charge q, it is simple to recover the electric field created by charge Q; simply divide by the magnitude of charge q ($\frac{kQ}{r^2}$). In a roundabout way, we have found a relationship between electric potential and electric field, using the relationships developed in earlier sections: take V, multiply by q, take the spatial derivative, divide by q. You may have noticed that this is overly complicated. Instead, we can simply take the derivative of the electric potential with respect to separation distance.

What does this mean? Refer to Appendix D or your favourite introductory calculus book if you need a refresher on calculus. The electric field is big

when the derivative of the electric potential is large, which occurs in regions where the electric field is changing rapidly. On a graph of electric potential as a function of separation distance, a large electric field corresponds to a steep slope. In the equipotential map representation, a large electric field corresponds to potentials that are close together (in those locations, the potential changes rapidly over short distances).

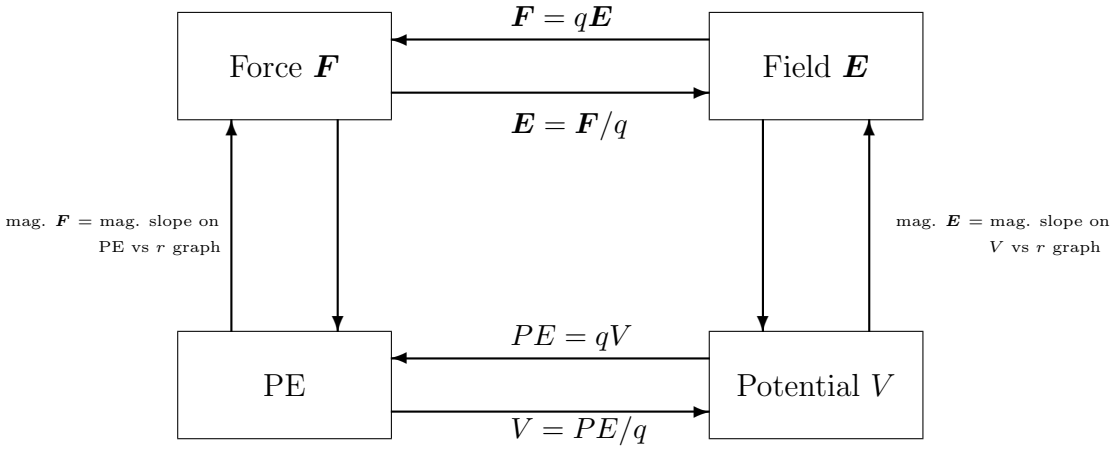
Four box diagram

In section 9-1-6 we first introduced all of the relationships between $\boldsymbol{F}$, $\boldsymbol{E}$, V, and PE in a diagram. This diagram is reprinted below. Study the relationships again, after reading more about electrical phenomena and again after completing DL activities on the subject matter. Memorising this diagram will not enable you to understand the physics in this section, nor to ace your quizzes. Instead, look at the diagram as an organising structure. Ask yourself the following types of questions.

- What is common between the quantities on the bottom (PE and V)?
- What about the items on the left side ($\boldsymbol{F}_{\text{field on } q}$ and PE)?
- How do you get from a quantity on the left to a quantity on the right?
- In what ways can we represent each of the quantities (what types of diagrams or graphs?)
- In a diagram of one item (such as a graph of electric potential as a function of separation distance), how can we gain information about other quantities (such as PE or $\boldsymbol{E}$)?

Electric field/force/potential relationships

Potential exists for time-independent electric fields only



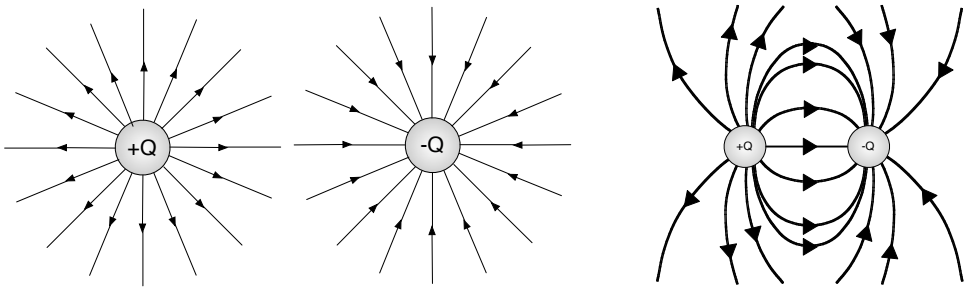
9-2-6 An in-depth example: the electric dipole

To make concrete all the ideas of electric phenomenon, we now work through the problem of an electric dipole. Note that this is one specific example of many available. It is more important to know the main ideas of this unit and how the ideas relate to one another (i.e. what is an electric field and how does it relate to electric potential?), than it is to know the specifics of the material presented in this section. The electric dipole was chosen because of its importance to magnetism.

The electric dipole consists of two charges of opposite signs separated by a small distance. Though this setup sounds fairly contrived, many molecules (water among them) act like dipoles, because within the molecule the charges separate leaving a net positive side and a net negative side. Much of the biology and chemistry you have studied has depended on dipole interactions. Note that there is no gravitational equivalent to a dipole, since there is no "negative" mass.

Let's start off simply. Imagine two charges, each of the same magnitude but opposite sign. In previous sections, we discussed the interaction between these charges; now we will discuss interactions between this *charge pair* and other charges. We are then interested in the field and potential created by *both* charges, not simply the field created by one charge at the location of the other charge.

We will start by considering the electric field. Field lines must begin on a positive charge and end on a negative charge (or continue forever into space).



(a) From single charges (b) Dipole field lines

Figure 9-2.7: Constructing field lines for an electric dipole. a) Fields lines from each independent charge. b) Field lines from entire dipole.

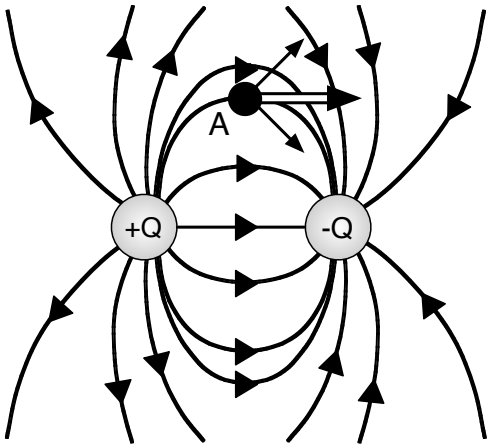


Figure 9-2.8: Sample field vectors at point A. Notice that the field vector is tangent to the field line, as expected.

The density of field lines represents the strength of the field. In this case, the charges have equal magnitude, so the density near each charge is the same. First consider the field lines from each charge in isolation, as shown in Figure 9-2.7a). Next we need to connect the lines in a sensible fashion. This is shown in Figure 9-2.7b).

We can check our work by calculating a few sample electric field vectors. Suppose the charges are oriented horizontally and separated by a distance L. Consider a point equidistant from both charges, such as the point labelled 'A' in figure 9-2.8.

We first draw the individual field vectors from the positive charge and the negative charge. For the positive charge, the vector points away from the charge. For the negative charge, the vector points toward the charge, at the same angle from horizontal. Point A is equidistant from each charge, so the magnitudes of the vectors are identical.

Using head-to-tail vector addition, add the vectors. The up and down components of the vectors cancel, leaving the sum pointing directly to the right. The sum, shown with a double arrow, points in the direction of the electric field at this point. Superimposing the field vector on top of the field lines, we see that the electric field vector is tangent to the electric field line at point A. This is reassuring, as field vectors are supposed to be tangent to field lines at every point.

Test yourself:
Check that the field vector is tangent to the field line at another point.

Next, we consider the electric potential. We wish to consider points very far from the dipole to have an electric potential of zero. Nearing the negative charge, the potential will get increasingly negative, and nearing the positive charge, the potential will get increasingly positive.

To directly calculate the potential at a given point, we will use the equation:

$$\mathrm{V} = \frac{kQ}{r}$$

but we need to consider the contributions from each charge. At a point equidistant between the charges, the distance r is the same for both contributions, though the signs differ.

$$\mathrm{V} = \frac{kQ}{r} + \frac{k(-Q)}{r} = 0$$

Any point equidistant to each charge has a potential of 0.

We will now draw the equipotentials using the information determined above and what we know about equipotentials from earlier sections:

- The potential far from the charges is 0.
- The potential equidistant from the charges is 0.
- The potential near either charge is increasingly positive (or negative).

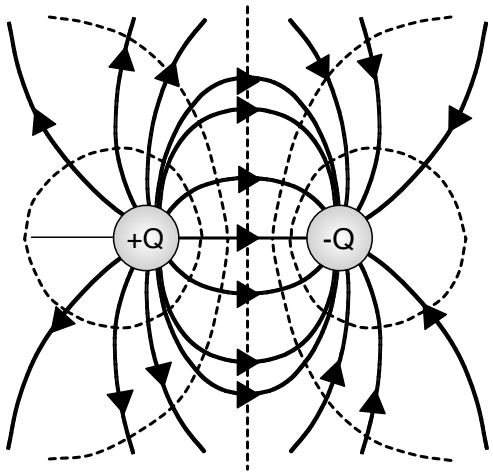


Figure 9-2.9: Field lines and equipotentials from an electric dipole.

– Equipotentials are perpendicular to field lines.

– In places where the field is strong, the potentials are close together.

Test yourself:
In §9-1-6-1, we summarised the field map, field lines, and equipotential representations in tabular form. Using the table, and the image above 9-2.9, make certain you understand each representation.

You may wonder why so much of 9-2 has been dedicated to the electric dipole. There are several reasons for this attention. For one, it has given us a chance to put together electric fields and potentials in a more complicated example. Additionally, dipoles are ubiquitous in chemistry, as you may recall from dipole-dipole bonding, for instance. Finally, as you will see in 9-3, the dipole field is of critical importance to magnetism.

9-2-7 Summary

1. Electric interactions involve electric charge, measured in units of Coulombs.
2. The electric field is a vector field. It contains all of the information required to determine the force on an object placed in the electric field,

using

$$\boldsymbol{F} = q\boldsymbol{E}.$$

Units of the electric field are J/C, or equivalently N/m.

3. By convention, electric fields point away from positive charges and toward negative charges. Field lines may only start or end on charges; if they do not end they either loop around on themselves or extend toward infinity.

4. The electric potential, V, is a scalar field, having units of Volts. How quickly the electric potential changes along a distance Δx indicates the *strength* of the electric field *in that direction.*

$$E = \frac{\Delta V}{\Delta x}.$$

5. Both the electric field and the electric potential can be determined entirely from the source charges. The principle of superposition allows us to add the effects of multiple charges to find a single, total E and V.

6. An electrical force is an interaction between two (or more) charges. For two point charges this charge can be calculated directly

$$|\boldsymbol{F}_{\text{point } q_1 \text{ on point } q_2}| = \frac{kqQ}{r^2},$$

or determined from the field created by the source charges, or found by seeing how the potential energy changes with distance

$$|\boldsymbol{F}| = \frac{\Delta PE}{\Delta x}$$

7. The concepts in this chapter have several common representations:

   – A vector map: This representation for the electric field shows individual field vectors at representative points.

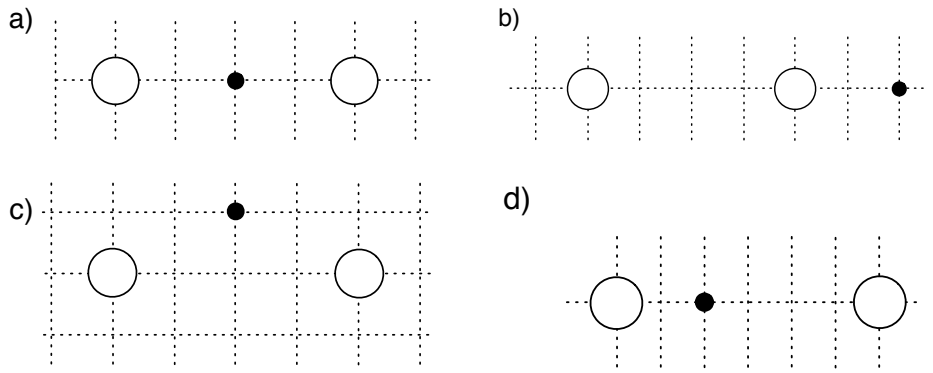
   – A field line map: This representation for the electric field uses continuous lines. The actual field is tangent to the field lines at any point. The strength is determined by how close together the field lines are.

   – Equipotentials: This representation for the electric potential indicates spots with equivalent potential. Where the potential changes quickly, the electric field is large.

9-2-8 Exercises

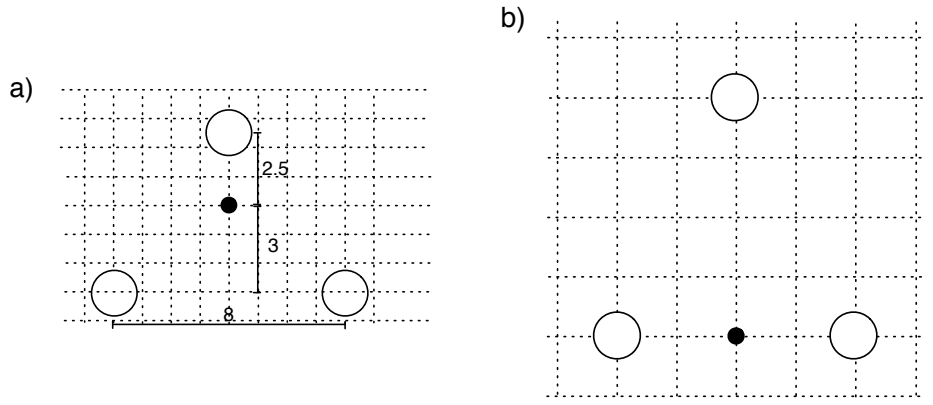
Questions 1-3:
In the following figures, the hollow circles represent the location of sources charges. The small black dots indicate the location of the test charge. In each problem, you are asked to reach some goal (like $\boldsymbol{E} = 0$). If the goal is *not* possible, explain *why* it is impossible. If the goal *is* possible, or determine the source configuration required to meet the goal. A single plus sign + indicates one unit of positive charge, and a single − sign indicates one unit of negative charge. You are asked to place as many + or − as necessary at the source charges to satisfy the requirements of the problem.

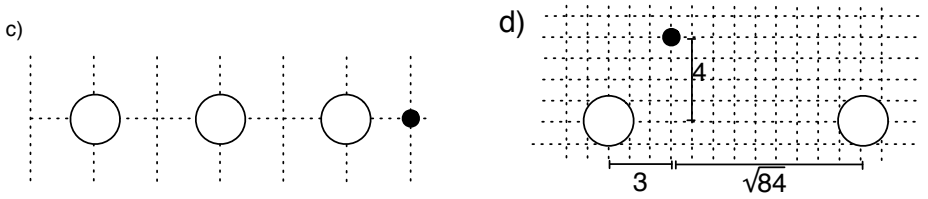
1. Goal: Electric field at the test charge location 0. In each case, add many + or − as necessary, or explain why the goal is not possible.



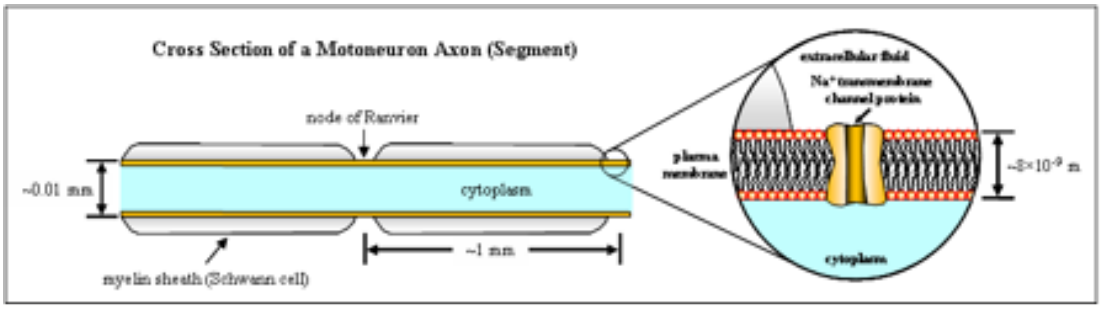
2. Goal: In each of the *preceding* examples, add as many + or − charges to the balls to make the net field point upward at the test location.

3. Goal: Electric field at the test charge location is 0.





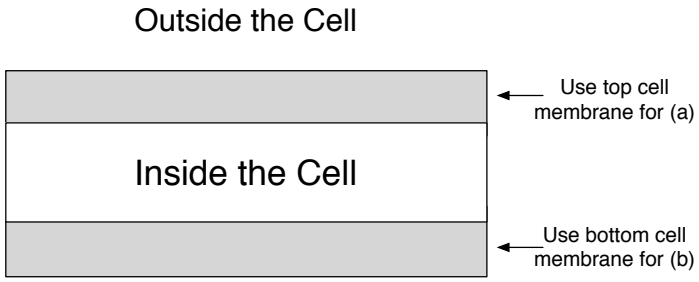
Questions 4-8 Human nerve cells (neurons) transmit messages by sending electrical impulses. The cell membrane contains channels that allow various ions, like Na^+, K^+, and Cl^- to selectively move in or out of the cell, thereby changing the charge distribution and the electrical potential. The voltage inside the cell differs from the voltage outside the cell across the length of the cell membrane. In this problem, we will model the electric field as uniform, meaning that the *magnitude* and *direction* of the electric field is the same throughout the membrane.



4. Before the cell fires, the ion distribution is such that the outside of the cell has a higher potential than the inside of the cell. This changes when the cell fires: at the peak of the nerve impulse, the cell membrane allows a rapid influx of Na^+ ions, giving the inside of the cell a net positive charge compared to the outside.

   The image shown below is a crude cross-section of the length of the cell, showing (from top to bottom) the outside of the cell, the top cell membrane, the inside of the cell, the bottom cell membrane, and again the outside of the cell.

   (a) Use a *field line map* to represent the electric field within the top membrane. Recall the the field is *uniform*. You may find it helpful to draw the charge distribution within and without the cell.

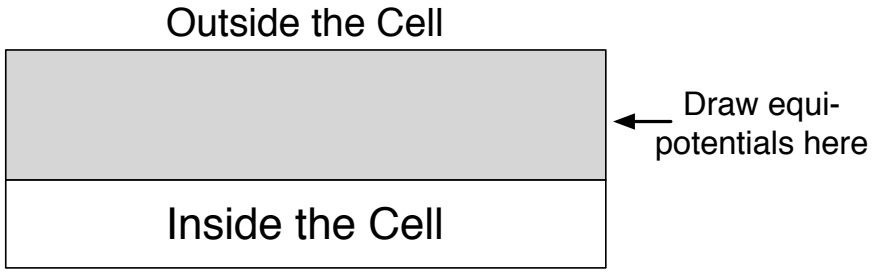


(b) Use a *vector map* to represent the electric field within the bottom membrane.

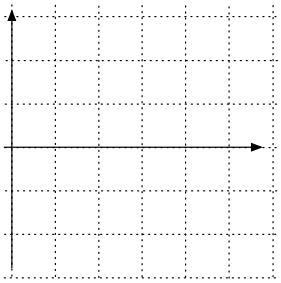
5. During the nerve impulse the voltage difference between the inside and outside of the cell is approximately 40 mV. If the cell membrane is 8×10^{-9} m thick, the cell diameter is 2×10^{-6} m, and cell length is 10cm, determine the strength of the electric field you drew in question 4.

6. Next, represent the voltage in two ways.

(a) Draw equipotentials within the membrane in the figure shown below. Space them 20mV apart.



(b) Graph the voltage as a function of position. Explain any assumptions you make in a complete sentence. Place y=0 at the inside border of the cell and membrane, and let y be positive as you move toward the outside of the graph. Be sure to label both axes with numerical values.



7. Explain how your answers to both a) and b) of question 6 reflect the electric field being uniform.

8. As mentioned above, the impulse occurs because sodium ions flow into the cell. Think about the very last sodium ion to flow into the cell. How does this cell's potential energy change as it passes into the cell (magnitude and sign)? Which direction is the force on the cell? (If you're curious about how the sodium ion manages to flow into the cell, that's good! Cellular regulation of this process is a balance of potassium leak channels, sodium-potassium pumps, open cells allowing diffusion, and voltage-gated sodium channels. A course including neurobiology, like NPB 101, will help you sort it out!)

Unit 9:

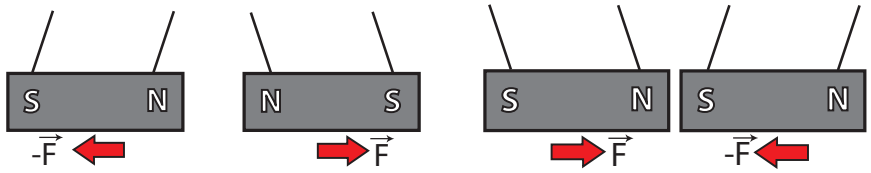
9-3: Magnetic fields

9-3-1 Magnetism and the B field

9-3-1-1 Magnets, north and south poles.

When we are given a set of magnets, we usually begin exploring their behavior by playing a bit with them. One of the first things we find is that we can get a pair of magnets to stick when some ends face each other, but turning one of the magnets around it is hard to get them to stick. We find that bar magnets can attract or repel other magnets depending on which ends are brought together. We need a way to label the two ends. The convention is that these two distinct ends of a magnet are called the *north pole* and the *south pole*. This convention is related to the use of magnets in navigation. The development of the navigational compass (around the 12th century) makes use of the interaction between bar magnets and the Earth, which for the purposes of magnetism is one large bar magnet. We can then use the Earth as our bar magnet of reference and use it to define the pole of all magnets. The poles are defined thus: If we suspend any magnet freely on a table (by a string or sticking it to a cork which is free to float on a liquid, to name some examples) the end of the magnet that points towards the north geographic pole of the Earth is referred to as the "north-seeking pole" or just the north pole. The opposite end is the "south-seeking pole" or just the south pole.

With this definition, we can go back and explore the behavior of magnets in a similar fashion as one can explore the behavior of electricity once the charge-sign convention has been established. If we bring together two magnets and want to make their north poles touch, we notice that there is a repulsion between the two magnets. If we flip one of the magnets so that now we make a south pole approach a north pole, we notice that the magnets will attract, even stick together. Therefore, we can conclude from these observation that



(a) Magnetic Repulsion (b) Magnetic Attraction

Figure 9-3.1: The attraction and repulsion of magnets. a) Similar poles repel each other (North-North in this case). b) Opposite poles attract.

like poles repel and opposite poles attract. This behavior is illustrated in figure 9-3.1, and is something you should also verify in your D/L. In figure 9-3.1 (a), we suspend two bar magnets from a pair of strings, such that the North poles of both magnets are in close proximity. The two bar magnets will exert a repulsive force on each other, represented by the arrows in the Figure. The repulsive force that each bar magnet exerts on the other will be equal in magnitude and opposite in direction (Newton's Third Law). In Figure 9-3.1 (b), we do the same thing but flip one of the magnets. Now, the North pole on the left magnet is close to the South pole on the right magnet, so the two bar magnets experience an attractive force.

One interesting aspect of magnets, a feature that is distinct from electric phenomena, is that magnets *always* have two poles. Imagine that you have a long bar magnet with the north pole on the right and the south pole on the left end. Let's say you want to break it into two pieces. One might think that you end up with only a north pole in the right piece and only a south pole in the left piece. It turns out that upon breaking it, there will be two new poles appearing right at broken ends, so each new piece will still have a north and a south pole, as in Figure 9-3.2.

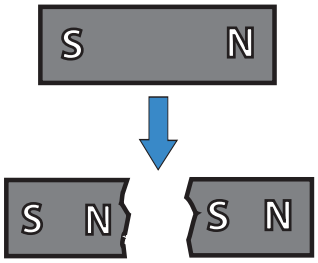


Figure 9-3.2: Absence of magnetic monopoles. Upon breaking a bar magnet, we end up with two magnets that still have both North and South poles.

This behavior has no analog in electricity, where one can separate positive and negative charges, and each charge can exist in an object separately. In nature, as far as we know, there are no north poles without a south pole, nor south poles without a north pole. This is referred to as the absence of magnetic "monopoles", and is one of the most important findings in magnetism.

9-3-1-2 The $\boldsymbol{B}$ field.

Once we have defined a convention for the two types of magnetic poles, we are ready to define the magnetic vector field $\boldsymbol{B}$. Imagine placing several round magnetic compasses on a plain sheet of paper on a flat table. In the absence of any other magnets, the compasses will all align themselves in the same direction: their north poles will be pointing towards geographic North by our definition. So we rotate the compasses such that the label "N" points to the top of the page. Assume that we now place a large bar magnet horizontally on the sheet of paper as shown in Figure 9-3.3, with the North pole on the right of the page. What happens? The magnetic needles of the compass, which are very light and can move freely (the equivalent of our test charges for the electric field) will reorient themselves in the presence of the bar magnet. The compass body, which is usually not magnetic, feels nothing and so remains with the "N" labels pointing to the top. The compass needles now do not point to the top; they will be pointing in different directions, as shown in the figure. The bar magnet produces a magnetic field in its surrounding space, and the magnetic needles of the compass reorient themselves because the field exerts a force on them. For example, the compass needle that is just right of the bar magnet is closest to the North pole of the bar magnet. Therefore the North pole of the compass needle will be repelled by the North pole of the bar magnet, so it will now point to the right. Examining the other compass needles, we can use this situation to define the direction of the magnetic field vector $\boldsymbol{B}$: at a given location in space, $\boldsymbol{B}$ will point in the same direction that a north-seeking magnetic compass would.

Another way to say this is to use similar language to the $\boldsymbol{E}$ field definition. Recall that we defined the direction of the $\boldsymbol{E}$ field as the direction of the force that a positive charge would feel at a given location in space. Similarly, the $\boldsymbol{B}$ field direction is defined as the direction of the force that a north-seeking pole would feel at a given location in space. So if we place a compass near the north pole of the bar magnet, the north pole of the compass will be repelled by the north pole of the bar magnet (and the south pole will be attracted to it) causing our compass to point as shown in Figure 9-3.3. Imagine now placing compasses one after the other, such that their compass needles are touching head to tail. If you follow this procedure starting near the north

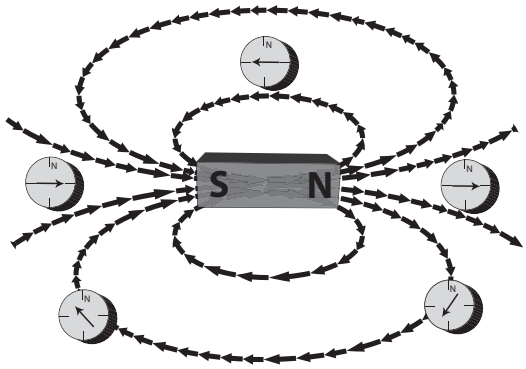


Figure 9-3.3: Magnetic field direction. Note that the $\boldsymbol{B}$ field points away from North poles and toward South poles.

pole of the bar magnet, you will find that you will end up eventually at the south pole, as shown by the progression of arrows in the figure. This schematic illustrates the idea that we can follow a line from North to South pole of the magnet. Just as we did with the $\boldsymbol{E}$ field, we can represent the magnetic field of a magnet by drawing the magnetic field lines, which will follow the path of the head-to-tail compasses in our example. Note that in the figure, there is one north pole and one south pole present, so we have two opposite type poles. This configuration is very similar to the electric dipole, where we had one positive and one negative charge present. You will notice that these two field configurations have a similar shape for the field lines.

9-3-2 Magnetic forces

9-3-2-1 Magnetic force on a moving charge

Until now we have talked about our everyday experience of magnetic fields originating from what are called "permanent magnets". But magnetic fields and electric charges are intimately linked, as we will soon see in greater detail. For now we will switch gears and consider how a charged particle behaves in a magnetic field, and in particular what force it feels. We do not really care *how* the magnetic field got there just yet, and for simplicity's sake we will imagine that the field was created by a magnet. This ignores the interesting question of what makes a magnet, a question we will return to in section 9-3-3. That is, we are going to start with the *field model* of magnetic fields:

$$\text{Unknown} \overset{\text{creates field}}{\Longrightarrow} \boldsymbol{B} \overset{\text{exerts force on}}{\Longrightarrow} \text{Charge } q$$

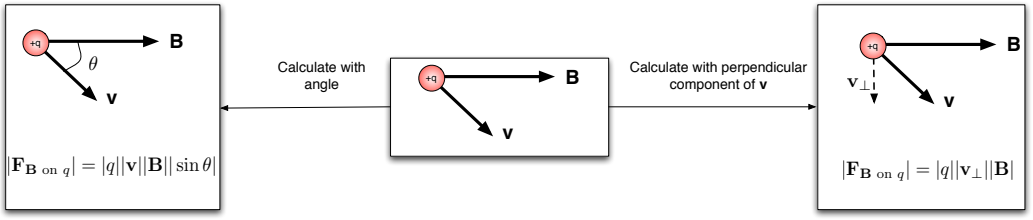


Figure 9-3.4: Showing two equivalent ways of calculating the magnitude of the magnetic force on a test charge

Magnetic forces are of course vectors with both magnitude and direction. We will begin by analyzing their magnitude, and leave the issue of direction to section 9-3-2-2.

Experiments demonstrate that the magnetic force exerted on a charged particle depends on its velocity. If a particle is not moving, it feels no force from the magnetic field! The magnitude of the magnetic field on a charge q travelling with velocity $\boldsymbol{v}$ is given by

$$\left|\boldsymbol{F_B}_{\text{ on } q}\right| = |q||\boldsymbol{v}||\boldsymbol{B}||\sin\theta|$$

where θ is the angle between the velocity $\boldsymbol{v}$ and the magnetic field $\boldsymbol{B}$.

Notice that if the velocity $\boldsymbol{v}$ and magnetic field $\boldsymbol{B}$ point in either the same or opposite directions then the force from the magnetic field is zero. In fact, to calculate the magnetic force all we need to know is the component of velocity perpendicular to the $\boldsymbol{B}$ field. This is what the $\sin\theta$ in the above formula does. Another way of rewriting the force from the magnetic field is

$$\left|\boldsymbol{F_B}_{\text{ on } q}\right| = |q||\boldsymbol{v}_\perp||\boldsymbol{B}|.$$

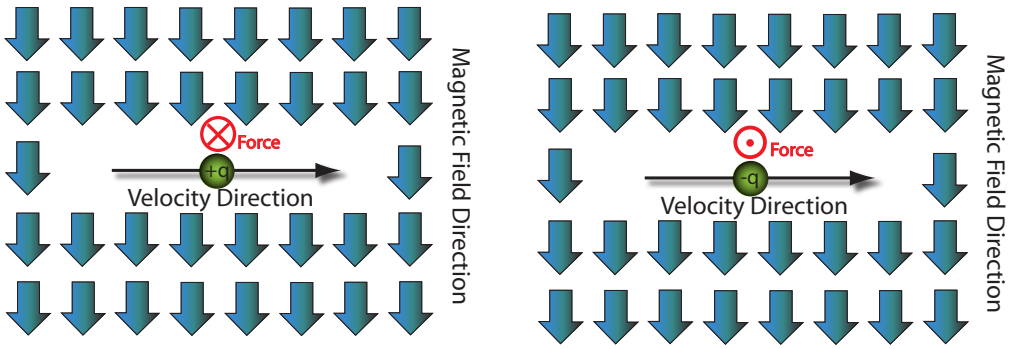
This is very similar to calculations of the magnitude of torque from 7B, in which only the component of force perpendicular to the lever arm mattered. Both approaches for calculating magnetic force are illustrated in figure 9-3.4

9-3-2-2 Right hand rule

When we looked at gravitational fields, we learned that the gravitational force was always in the same direction as the field. When looking at the electric field, we learned that the force was in the same direction as the field

(for a positive charge) and in the opposite direction to the field (for a negative charge). The magnetic field is quite different – the force on a charged particle *never* points in the direction of the magnetic field.

We noted that the velocity $\boldsymbol{v}$ of the charged particle determined the magnitude of the magnetic force. It is also needed to determine the *direction* of the magnetic force. Experimentally we find that the magnetic force on a test charge is at 90° to the magnetic field $\boldsymbol{B}$, and 90° to the velocity of the test charge $\boldsymbol{v}$. If the magnetic field $\boldsymbol{B}$ and velocity $\boldsymbol{v}$ are not pointing in the same direction, then there are only two possible directions that are at 90° to both of these directions. One of these directions is the direction of the magnetic force on a positively charged particle; the other is the direction of the force on a negatively charged particle. To correctly pick the direction for



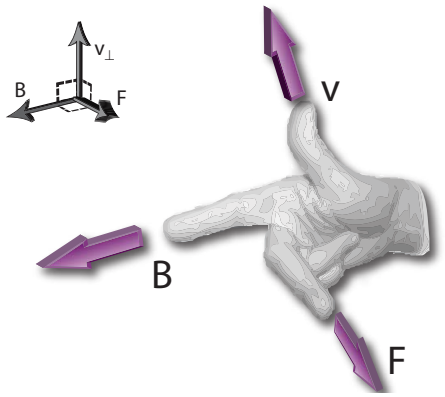
(a) Force on Positive q.

(b) Force on Negative q.

Figure 9-3.5: A charge moving with velocity $\boldsymbol{v}$ in the presence of a $\boldsymbol{B}$ field will experience a force $\boldsymbol{F}$ perpendicular to both $\boldsymbol{v}$ and $\boldsymbol{B}$. The sign of the charge will determine the direction of $\boldsymbol{F}$.

the positive particle, we use *right hand rule #2* (RHR #2). We point our right thumb in the direction the particle is going ($\boldsymbol{v}$), our right index finger in the direction that the $\boldsymbol{B}$ field is going, and then our middle finger on our right hand when bent will point in the direction of the magnetic force[1]. The mnemonic "**v**ery **b**ad **f**inger" may help you remember the order. The diagram below may help:

[1]If you are double jointed, just pretend that you are not for the purposes of using the right-hand rules.



To find the direction of magnetic force on a negatively charged particle, find the direction of the magnetic force on a positively charged particle and reverse it.

All of the above was done under the assumption that the magnetic field $\boldsymbol{B}$ and the velocity $\boldsymbol{v}$ were in different directions. How do we determine the direction if we decide to fire a charge in *exactly* the same direction as the magnetic field? The answer is it does not matter – if the velocity and magnetic field are in exactly the same direction then the magnitude of the magnetic force is zero! As we saw before, if the $\boldsymbol{B}$ field and the velocity $\boldsymbol{v}$ are parallel (or anti-parallel), the angle θ between them is zero; therefore $\sin\theta$ is also zero, and the force will be zero.

With both the direction and the magnitude determined, we can now summarize the magnetic force on a test charge q traveling with speed v:

$$F_{\boldsymbol{B}\text{ on }q} = \begin{cases} \text{Magnitude} = |q||\boldsymbol{v}||\boldsymbol{B}||\sin\theta| = |q||\boldsymbol{v}_\perp||\boldsymbol{B}| \\ \text{Direction} = \text{middle finger in RHR\#2 if } q \text{ is positive} \\ \qquad\qquad\quad \text{opposite middle finger in RHR\#2 if } q \text{ is negative} \end{cases}$$

9-3-2-3 Magnetic force on a current-carrying wire

Regardless of whether a charge is in a vacuum or inside a material, it will experience a force when it moves across magnetic field lines. Therefore, a wire with a current will experience a force that is the vector sum of all the forces acting on the charge carriers that are individually moving in the wire creating the current.

Consider a straight wire segment of length L with a current I flowing from left to right, placed on the page. Imagine that there is a $\boldsymbol{B}$ field in this region that makes an angle θ with respect to the wire. If the charges

on the wire are moving at an average speed v, the time they need to travel the length L is $\Delta t = L/v$. The amount of charge that flows in this time is $q = I\Delta t = IL/v$. Therefore the force exerted on the wire is

$$F = qvB\sin\theta = (\frac{IL}{v})vB\sin\theta = ILB\sin\theta$$

The direction of the magnetic force on a wire is also given by the same right-hand rule used for single charges. Of course, we now know that it is electrons which are actually moving in the wire, not positive charges as is the convention for the current I, but the electrons are moving in the opposite direction as the conventional current I, so in both cases, the direction of the force obtained by applying the right-hand rule is exactly the same, as it should be.

9-3-3 Magnetism and currents

9-3-3-1 What creates magnetic fields?

In the last section we learned that a magnetic field affects *moving charges*. By Newton's third law, the moving charges must exert an equal and opposite force on whatever produced the $\boldsymbol{B}$ field. That is, the moving charge must create its own $\boldsymbol{B}$ field! By using Newton's third law we can complete the description of the indirect model we started above:

$$\text{Moving charges} \overset{\text{creates field}}{\Longrightarrow} \boldsymbol{B} \overset{\text{exerts force on}}{\Longrightarrow} \text{Moving charge } q.$$

As we learned in Physics 7B we can think of moving charges as a current; a concept that is particular useful if we have a steady flow of charge. So the indirect model becomes:

$$\text{Current} \overset{\text{creates field}}{\Longrightarrow} \boldsymbol{B} \overset{\text{exerts force on}}{\Longrightarrow} \text{Moving charge } q.$$

While we have worked this out, it is far from clear what currents and moving charges have to do with anything related to the fridge magnets or bar magnets that make magnetism familiar to us. In essence we have cheated: the ideas of how a magnetic field affects moving charges were not known until the mid-1800s. Before that, the only things known about magnetism were some materials can produce magnetic fields and these attract (or repel) certain ends of other similar materials, and that the Earth had its own magnetic field which aligns these magnetic materials. These facts were known to the

Greeks as early as 600 BC. The question of why certain materials where magnetic while others did not appear to be, and what phenomenon created these magnetic fields was not addressed until 1820.

In 1820, Dutch physicist Hans Christian Ørsted had set up an experiment to show that large electric currents could be used to heat a wire. While demonstrating this to a group of students in his house, he noticed that a compass on his bookshelf changed direction whenever his "kettle" was switched on. After months of investigation, Ørsted concluded that an electric current could create a magnetic field. This was big news at the time, because prior to this only magnetic fields were known to affect other magnetic materials. This was a watershed moment in the history of science, as it was the first link between electric and magnetic phenomena. Originally we experience these as two distinct forces, two distinct fields. Ørsted's finding was the first step on the road that led humankind to find that these apparently dissimilar phenomena were in fact linked. This unification of seemingly disconnected ideas is still at the core of fundamental research: much hope is placed on possibly unifying *all* forces in nature.

The finding that electricity and magnetism are linked caused a huge revolution in science, but we now want to return to our question of what makes a magnet a magnet. Ørsted showed us that electric currents created a magnetic field, but where are the currents in a magnetised piece of Iron? People could not answer this question until the late 1800s, and even then they were met with skepticism. The answer relied on the existence of atoms: in a nutshell, the origins of magnetism are found in the electric currents produced by the electrons orbiting (i.e. making a current loop) atomic nuclei, as well as their spin orientations. Spin is a purely quantum mechanical phenomenon, but for the purposes of thinking about magnetism in our current discussion, consider spin is an additional way to produce a "current loop" (albeit a loop with the smallest size you can think of!) by endowing the electrons with an intrinsic angular momentum. This "intrinsic spin" can only have two possible values. From chemistry you would have heard that the electron is "spin up" or "spin down". These also play a role in magnetism. The fact that atoms exist is something we now take for granted, but which most scientists thought of as ludicrous until 1905 (due in large part to a separate contribution by Einstein)! After all, they reasoned, every other thing they knew in the world lost energy due to friction, and the idea of atoms with electrons that just kept going around perpetually was absurd.

To summarize, all our experiments point to the following finding: to get a magnetic field, we have to have a net current. How does this idea explain magnetism in materials? Imagine helium, an atom with two electrons. Now if these electrons go around in opposite directions, the currents they produce will be opposite to each other. The magnetic field of one current loop will cancel the magnetic field of the other, leading to no net magnetic field. The spin also affects the magnetic field, and if the spins are pointed in the same direction (both up or both down) the field gets stronger, while if they are aligned in opposite directions the spin gets weaker. In helium, the spins of the two electrons are paired up-down, so helium would not be very magnetic. In fact, helium is an "anti-magnet" and tries to stop any magnetic field going through it. This effect is called diamagnetism and is a manifestation of Lenz's law which we have not covered yet.

There are a lot of atoms that have an odd number of electrons that are not very magnetic. So far we have asked whether each individual atom will be influenced by a magnetic field. The missing ingredient you need is for the different atoms to strongly interact with one another. Since a material is made up of $\sim 10^{23}$ atoms, tiny effects can become large if each atom contributes a little magnetic field, and if each atom contributes a magnetic field *in the same direction as its neighbour*. Otherwise one atom may decide to have a net current clockwise, and its neighbour counter- clockwise and the magnetic fields will almost cancel. The materials that have their outer electrons interact strongly are metals (they have almost "free" outer electrons) so highly magnetic materials are generally metals.

At the end of all this, we see that magnetism boils down to moving charges affecting moving charges. It is no surprise that an explanation of bar magnets took so long, as it required a serendipitous observation of how two superficially unlike phenomena (electricity and magnetism) affected one another and a model of the atom. We have already studied how a magnetic field affects a moving charge (see 9-3-2-1) Now we turn to the quantitative question of how, exactly, a current produces a magnetic field.

9-3-3-2 Ampère's Law

We will first study a simple test case: a long straight wire carrying a current. We want to understand the magnetic field produced by this wire, i.e. how strong it is in magnitude, where does it point to (recall it is a vector), and how does it vary from point to point. In other words, we want to map the

B Field produced by a long straight wire.

Field from a long, straight wire.

We will retrace some of Ørsted's steps. He showed that a current-carrying conductor produces a magnetic field. A simple way to demonstrate this is to place several compass needles in a horizontal plane (for example, the surface of a table) near a long wire placed vertically, i.e. perpendicular to the surface of the table. Let's assume the current direction to be coming from the bottom and going toward the top of the table. We hook up the wire to a battery so that we can turn the current on or off. When there is no current in the wire, all needles point in the same direction (magnetic north). As soon as we turn on a circuit and current starts flowing in the wire, all needles will deflect. We have produced a magnetic field! The first thing that one notices when doing this experiment, is that the needles orient themselves in a definite pattern. If we draw a circle on the table with the wire at the center and we place the compasses along the circle we just drew, we will notice that all the compass needles will orient themselves *tangential to the circle* we just drew. In other words, the ***B*** field lines for the long straight wire at a distance r from the wire will have the shape of concentric circles of radius r. For our experiment with the current coming out of the table, we find that the ***B*** field direction is counterclockwise along the circles. If we flip the direction of the current, the compass needles all deflect by 180°, i.e. they will still point tangent to the circle, but now their North poles point in a clockwise direction. This observation for the direction of the ***B*** field can be summarized with the following convenient rule, Right-hand-Rule #1: :

Point the thumb of your right hand along a wire in the direction of the conventional (positive) current. Your fingers will now curve naturally in the direction of the magnetic field.

What can we infer about the magnitude of the magnetic field? By symmetry, the magnitude of the magnetic field should be the same everywhere along the circle. Why is this? Nothing is special about any particular point along the circle; they are all equidistant from the wire. Likewise, if the wire is infinitely long, there is nothing special about where we chose to place the horizontal plane where we put our compass magnets. Moving this horizontal plane up or down along the wire, if the wire is infinitely long, should also have no effect in our results. To pick a coordinate system, we can set the plane of the wire to be the x-y plane, and the direction along the wire to be the z-axis. By symmetry, the magnetic field can therefore not depend on the z coordinate. Therefore, we expect that the magnitude of the field at any point will depend only on the perpendicular distance between the wire and

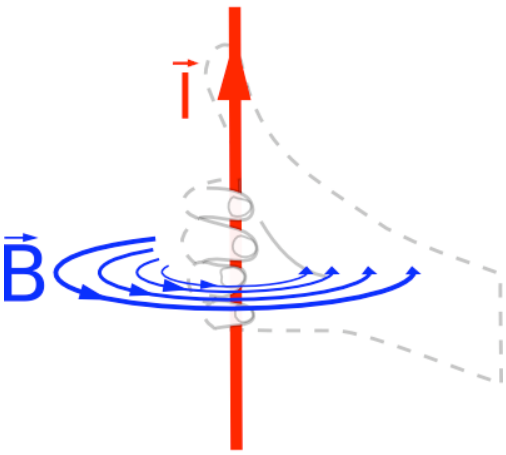


Figure 9-3.6: Right hand rule for currents

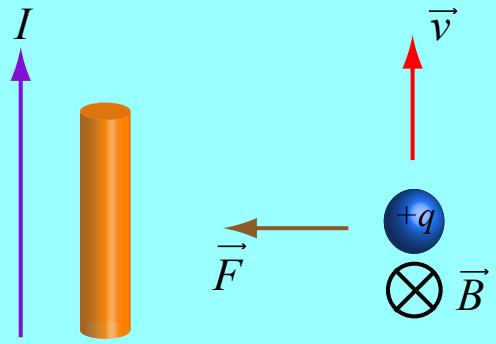
that point. All points at the same distance r will have the same magnitude (that is why the shape of a circle of radius r around the wire comes naturally). It is also not unexpected that the magnitude of the field will be larger if we have a larger current. It turns out that the magnitude is proportional to the current. Likewise, you probably expect that if we start moving away from the wire, the magnetic field will get weaker the farther we move from the wire. The equation that relates all these quantities to the magnetic field magnitude at a point $\vec{r}$ is:

$$|\boldsymbol{B}(\vec{r})| = \frac{\mu_0 I}{2\pi r}$$

where the I is the current on the wire and r is the perpendicular distance from the wire to the point we are interested in. In our experiment with a wire on a table where we set the x-y plane to be the plane of the table, the distance r would be the length of the 2-D vector that points from the wire to the point. The constant μ_0 is a proportionality constant called the magnetic permeability of the vacuum, which has the value $\mu_0 = 4\pi \times 10^{-7}$ T $\cdot$ m/A.

Example #1:

A long straight wire carrying a current I produces a magnetic field of 1.0×10^{-4} T at a distance of 2 cm away from the wire. Find the current I carried by the wire. How close to the wire is the field strength 1T? A proton is moving at $v = 1.5 \times 10^3$ m/s parallel to the wire and in the same direction as the current a distance of 1.0 cm away from the wire. Find the magnitude and direction of the magnetic force exerted on the proton by the magnetic field produced by the wire.



Solution:

We use Eq. 9-3-3-2 to solve for the current I. Once we find the current, we set the field equal to 1 T to solve for the distance r from the wire.

Solve for I: $I = 2\pi r|\boldsymbol{B}|/\mu_0$

Insert numerical values: $I = 2\pi(0.02\text{m})(1.0 \times 10^{-4}\text{T})/4\pi \times 10^{-7}\text{Tm/A} =$ 10 A

Solve for r with new value for $|\boldsymbol{B}|$: $r = \mu_0 I/2\pi|\boldsymbol{B}| = 2\ \mu\text{m}$

Direction of $\boldsymbol{B}$: By RHR1, the magnetic field produced by the wire at the proton's location is going into the page.

Direction of $\boldsymbol{F}$: By RHR2, the magnetic force on the proton is directed towards the wire.

Magnitude of $\boldsymbol{F}$: We know that $|\boldsymbol{B}|$=0.1 T at 2 cm from the wire. At 1 cm from the wire, since $|\boldsymbol{B}| \propto 1/r$, we must have $|\boldsymbol{B}|$=0.2 T. We now insert —$\boldsymbol{B}$— into the equation for the magnetic force, $q|v||\boldsymbol{B}|\sin\theta =$ $(1.6 \times 10^{-19}\text{C})(1.5 \times 10^3\text{m/s})(0.2\text{ T})(\sin 90°) = 4.8 \times 10^{-17}\text{N}$

Field from a current-carrying coil

We would now like to describe the magnetic field from another simple configuration of electrical current. Consider a coil of radius r_0, made up of N

loops, carrying current I, as shown in Figure 9-3-3-2

Suppose we are interested in the magnetic field along the coil's axis, at point P, a distance z along the axis from the coil. We can think of the coil as N copies of a single loop of current I. Each loop is just a curved wire of current. If we apply the right-hand-rule to a small piece of one of these loops, we find that no matter which part of the loop we choose, the resulting $\boldsymbol{B}$-field is pointed up for our point P, or indeed any point on the axis of the coil. It makes sense that the further from the coil we get, the smaller the magnitude of $\boldsymbol{B}$ will be, and that it will be directly proportional to the number of loops we have as well as the current flowing through the loop as before.

Indeed, it turns out the magnetic field at distance z along the axis (as long as we are relatively far from the coil) is given by:

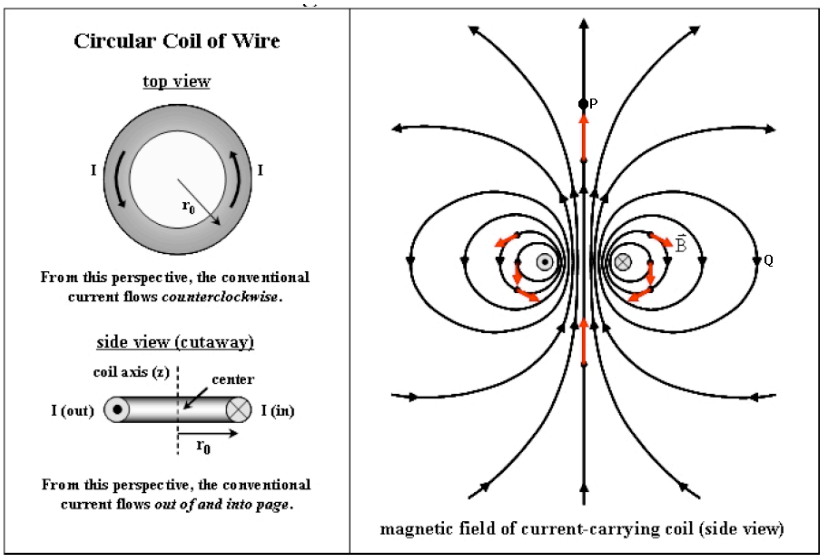
$$|\boldsymbol{B}| = \frac{\mu_0 N I r_0^2}{2z^3}$$

Now, what if we're instead interested in the magnetic field at point Q, in the plane of the coil? Again we turn to the right-hand-rule for currents. Notice that this time not all parts of each loop result in a contribution to the $\boldsymbol{B}$-field in the same direction. The side of the loop closest to Q gives a contribution to the field in the *down* direction, while the side farthest from Q contributes in the *up* direction. Since the field strength drops with distance, the closest side has the larger contribution, and the overall field is in the *down* direction. But one consequence is that the magnitude of the field at Q is significantly less than the magnitude of $\boldsymbol{B}$ on the axis at a similar distance, since the field on the axis has none of this cancellation.

The overall magnetic field lines are shown in the figure. Note the similarity with the field lines of a bar magnet from Figure 9-3.3. The similarity demonstrates that the magnetic field of a permanent magnet really can be described as the result of many small current loops lined up with each other.

9-3-4 Magnetic induction

We have seen how Ørsted was able to demonstrate that electric currents can produce magnetic fields. The English physicist Michael Faraday, a brilliant experimentalist, was the first to demonstrate the converse effect: magnetic fields can be used to produce electric currents. This is now called the principle of magnetic induction. It is interesting to note that Faraday had little formal schooling, so mathematics was by no means his strength. Nevertheless, he was one of the most influential scientists not just of his time, but his



contributions continue to find applications to this day. For example, when he demonstrated that magnetic fields could be used to produce a current in a wire loop, politicians were not impressed as they failed to see the use of it. It turns out that this is the critical step in making power plants and making electricity available outside of the techniques of kite-flying-in-a-lightning-storm and carrying large arrays of batteries. The alternating-current circuits that power all the electrical grids of the world have as part of their components a generator that is based on magnetic induction. More recently, highly fuel-efficient vehicles such as the gas-electric hybrid cars employ a technology called regenerative braking. This uses a device that can give power to the wheels of the car by means of an electric battery, and can recharge the battery during braking by running the circuit in "reverse", transforming the kinetic energy of the car's motion into electric potential energy stored in the battery and saving fuel as a result. This recent application has large implications for the world's economy and is of global environmental impact, and at its core lies Faraday's principle of induction: that we can transform magnetic fields to electric currents.

9-3-4-1 Magnetic flux definition

Before we tackle the actual form of the principle of magnetic induction, we first need to define a quantity which is crucial to understand it quantitatively: the concept of magnetic flux. Let us discuss first the idea of flux in general using a familiar example: rain falling on the windshield of a car. Let us suppose that we want to quantitatively determine the amount of rain that

hits the windshield of the car. For simplicity, let us first assume that the rain is falling vertically down, and that the shape of the windshield is a rectangle. Let us further simplify by assuming you are in a parked car, i.e. it is not moving. If we want to find how much rain hits the windshield, we need to consider chiefly these three variables:

- The amount of rain.
- The size of the windshield.
- The orientation of the windshield relative to the rain.

Let's discuss each in turn. If it is raining hard, there will be a lot more raindrops hitting the windshield than if it is raining lightly. If the size of the windshield is large, likewise there will also be more raindrops hitting it. The orientation between the rain and the windshield will also determine how much rain hits the windshield. For rain falling vertically down, if the windshield was also vertical, there will be no rain hitting the windshield (in the idealization that the windshield is infinitely thin). Conversely, if the windshield was almost horizontal (or if we consider the sunroof of the car, which is completely horizontal) the amount of rain hitting it will be much larger. This idea of calculating the amount of rain hitting a surface can be generalized by the concept of *flux*. We can consider the flux of a vector field through a given surface area. One can think of a flux of "something" as wanting to find out how much of that "something" passes through a given area. In our example, rain falling vertically gives us the idea of a vector field. More rain means that the magnitude of the vector field increases. A larger windshield means a larger surface area. The orientation is given, by convention, between the direction of the vector field and the vector *normal to the surface area*. For the particular case of the magnetic field vector $\boldsymbol{B}$, we define the magnetic flux Φ through an area A as

$$\Phi = \boldsymbol{B} \cdot \boldsymbol{A} = |\boldsymbol{B}||\boldsymbol{A}| \cos\theta$$

where the angle θ is the angle between the magnetic field vector $\boldsymbol{B}$ and the vector normal to the surface area A. From our rain example, you can see that when the rain is falling vertically down, and the windshield surface is horizontal, the vector normal to the area will also be vertical. Hence, the angle in the equation will be $\theta = 0°$, and $\cos\theta = 1$ leading to a large flux. If the windshield surface is also vertical, the vector normal to the surface area will now be horizontal, the angle θ will be $90°$, so $\cos\theta = 0$ and the flux will vanish (no rain hits the windshield in our example, as we had surmised). Note that for an open surface such as a windshield we still have the freedom

to choose the normal vector on either side of the surface. This will have no physical effect, but will simply change the value of the angle θ by 180°, in other words, the flux as we defined it changes sign. We will see shortly that what will be important physically will be changes in flux, not so much the actual value of the flux. So any of the choices will lead to identical changes in flux, resolving any ambiguity. To summarize, the variables of interest when calculating the magnetic flux through an area will be:

- The magnitude of the $\boldsymbol{B}$ field.
- The magnitude of the area under study.
- The angle between the $\boldsymbol{B}$ Field vector and the vector normal to the area.

9-3-4-2 Faraday's law of induction

Faraday's Law Now that we have defined the magnetic flux Φ, we can describe Faraday's observations quantitatively. He sought out to describe a connection between the magnetic field and a current in a wire in the presence of the field. For a current to flow, we always have to provide it with a closed wire loop. The area we will consider in the definition of flux will be the area within the wire loop. Note that the wire can be looped in a circular, square, rectangular, triangular, or other shape. The flux under consideration is the magnetic flux through the area enclosed by the wire loop. Faraday then asked what happened if you placed a magnet close to the loop and let it sit there. Would a current appear in the presence of the magnet? He carried out the experiment and found that there was no current in the loop. However, if you move the magnet away, then for a brief instant a current appears. If you move it back, then a current appears. What Faraday found is that **there will be an induced current, and therefore an induced voltage, only when the magnetic flux changes over time**. We say that this is an *induced* current because there is no battery, no voltage source in the usual sense, to create the current. The current is *induced* in the wire by the magnetic field. He called the induced voltage the induced "electro-motive force", or induced EMF for short, denoted by $\mathcal{E}$, and you can still find it under this name in many text books. We therefore refer to his findings as Faraday's Law of Magnetic Induction. Specifically what he found was that

- The induced voltage $\mathcal{E}$ is proportional to the rate of change of the flux with time, $\Delta\Phi/\Delta t$.

- If you add loops to the wire coil, each loop will contribute equally. If you have N coils, the induced voltage will be N times as strong.

We now summarize these findings in the equation that embodies Faraday's Law:

$$\mathcal{E} = -N\frac{\Delta\Phi}{\Delta t}$$

What this means is that you need to have a changing magnetic flux to produce an induced voltage. If the magnetic flux does not change with time, then there will be no current. Only if the magnetic flux changes with time will we observe a current. Furthermore, the faster the flux changes, the larger the induced voltage. You can picture this last statement in the following way. If you are generating an induced current by moving a magnet close to a wire, you will measure a larger induced current in the loop if you move the magnet quickly than if you move it slowly. The magnitude of the rate of change is proportional to the voltage: the faster the magnetic field changes, the greater the induced current and induced voltage.

Note also that Faraday's law focuses only on the effect of a magnetic field on a wire. For simplicity, we discussed producing this magnetic field using a magnet. However, we also can produce the magnetic field by using a current in another wire. This is in fact how Faraday studied the induced current and induced voltages.

The equation of Faraday's Law also has a sign. What is the physical meaning of the sign? It has to do with the *direction* of the induced current in the loop. We have not said how we are to choose between the two possibilities for the direction of the current. We find that the way nature works is that the induced current flows in such a way that the magnetic field, and its magnetic flux through the area of the current loop, must be such that they *oppose the changing magnetic flux* that induced the current in the first place. This is known as Lenz's Law. We can illustrate how this works with some examples.

Consider a circular loop of wire on the plane of the page. There is no magnetic field in the region of the wire, and at $t = 0$ a magnetic field appears pointing straight out of the page that increases in magnitude linearly with time. From Faraday's Law, we know that there will be an induced voltage, because this meets all the requirements for a change in flux. The flux is zero at first because there is no $\boldsymbol{B}$ field. After $t = 0$ the flux is non-zero. The area of the circle (which is the area enclosed by the circular wire) is constant, the angle between the normal to the area of the circle and the $\boldsymbol{B}$ field is $\theta = 0$,

so $\cos\theta = 1$ and is constant, and the magnitude of the $\boldsymbol{B}$ field increases with time. It is the last part that contributes to the rate of change in flux, producing an induced voltage. The induced voltage will be zero for $t < 0$ and then will be *constant* for $t > 0$. Recall that if the magnitude increases linearly with time, the rate of change (the slope) of the $\boldsymbol{B}$ field magnitude is a constant, hence the induced voltage will be constant. Will the current flow clockwise or counterclockwise? Now we use Lenz's Law. With the normal chosen pointing out of the page, i.e. in the same direction as the $\boldsymbol{B}$ field, the flux is positive and increases with time. The induced current should *oppose* this. Hence, it should produce a magnetic flux going through the loop that should be negative. With the normal pointing out of the page, we see that the $\boldsymbol{B}$ field produced by the induced current, which we call the *induced field* ($\boldsymbol{B}_{\text{ind}}$), should be pointing into the page. By the RHR, if we curl our fingers in the direction of the induced field with our hand around the wire, our thumb should indicate the direction of the induced current. We then see that the induced current for this case is flowing in a clockwise direction. A similar analysis should be carried out for each case you encounter. You can try it by yourself and figure out the current direction in the following variations on the above example:

- No field before $t = 0$ and a $\boldsymbol{B}$ field pointing into the page for $t > 0$.
- A constant $\boldsymbol{B}$ field pointing out of the page before t=0 and decreasing linearly with time from $t > 0$.
- A $\boldsymbol{B}$ field that looks like a triangular wave pattern with period T: starting from zero magnitude at t=0 rising linearly with time up to a maximum magnitude B_{max} at $t = T/2$ and then decreasing linearly with time down to zero magnitude at $t = T$ and repeating itself over a period of T seconds.

Real world applications of Faraday's Law

We started discussing Faraday's law by considering moving a magnet near a loop of wire. We have found that this produces an induced current in the wire. This idea has found its way into many applications in today's society.

- Seismograph

  One way to exploit Faraday's Law is to realize that if we attach a magnet to anything that moves and place it near a loop of wire, any movement or oscillation in the object will be picked up as an induced current in the wire loop. We can thus translate physical movements

and oscillations into electrical impulses. In all devices of this kind, the movement or oscillation is measured between the position of a coil relative to a magnet, whose movement causes the current in the coil to vary, generating an electrical signal. For example, the vibrations on the earth produced by an earthquake produce a current that can be amplified to drive a plotting pen. This is how a seismograph operates.

- Guitar Pickup. Les Paul, a pioneer musician of pop-jazz guitar, applied Faraday's Law to the making of musical instruments and invented the first *electric guitar*. The "pickup" of an electric guitar consists of a permanent magnet with a coil of wire wrapped around it several times. The permanent magnet is placed very close to the metal guitar strings. The magnetic field of the permanent magnet causes a part of the metal string of the guitar to become magnetized. When one plucks the string, these vibrations of the magnetized string create a changing magnetic flux through the coil of wire surrounding the permanent magnet, which "picks up" the vibrations that generate an induced current in the coil which can then be sent to an amplifier, to the pleasure of rock fans everywhere.

- Electric Generator. An electric generator is used to efficiently convert mechanical energy to electrical energy. The mechanical energy can be provided by any number of means, such as falling water from a hydroelectric dam (where the water converts its potential energy into kinetic energy as it falls), expanding steam in a thermo-electric power plant (where the steam can be heated by burning coal or oil, or by a the heat released in a controlled nuclear reaction) or your hand twisting a lever in a circle. In all cases, the principle is the same, the mechanical energy is used to move a conducting wire coil, typically by rotating it, inside a magnetic field. In this case, the area of the coil is the constant, the magnitude of the field is constant, so the $\cos\theta$ term is the one that produces a changing flux, i.e. the change in the relative orientation between the $\boldsymbol{B}$ field and the normal to the area of the coil. For this case, consider the simple scenario where we rotate the coil with constant angular speed ω. The rotation angle is given by $\theta = \omega t$, and the flux will be proportional to $\cos\omega t$. From differential calculus, the time rate of change of the flux will then be proportional to $\omega \sin\omega t$, so the induced current will oscillate sinusoidally. In other words, the current in the coil alternates in direction, flowing in one direction part of a cycle and in the other direction for the second part of the cycle. For this reason, this is referred to an *alternating current generator*, or

simply an AC generator. The standard plugs you use to power all of your electrical appliances are all powered by an electric generator of this form.

- Electric Motor.
- Hybrid cars: regenerative braking.

Unit 9:

9-4: Electromagnetic waves: light

When we started discussing the electric and magnetic fields they seemed to be quite separate. We knew that any electric charge (moving or not) created an electric field, and that any *moving* charge created a magnetic field. Other than that there was not a lot of similarity: the forces produced by the electric field looked very different than the forces produced by the magnetic field, and there were not "magnetic charges" (monopoles) for magnetic field lines to start or end on.

Then we started to discuss induction in §9-3-4, where we discovered that changing the magnetic flux through the loop caused an electric current to flow. If we kept the magnetic field constant and moved the wire then we could make some sense of this: by pulling the wire we had moving charges and by calculating the force on the moving charges we could show that under some circumstances we got a current. These circumstances corresponded to precisely those situations where the magnetic flux inside the loop changed, and it was simply easier to use the magnetic flux definition.

But one detail has been "swept under the rug" in the preceding analysis. We also learned that changing the flux by *changing the magnetic field* also caused a change in flux, and hence a change in current. If we model our charges as starting at rest ($v = 0$) then the magnetic field (changing or not) does not seem capable of forcing them to move as $|\boldsymbol{F}_{\text{B on charge}}| = |q\boldsymbol{B}\boldsymbol{v}\sin\theta| = 0$. How does the current start? The answer to this question is that changing a magnetic field *produces* an electric field. The electric field so-produced does not begin or end on charges; instead it connects with itself so that the field lines don't start or end (similar to the magnetic field). We should emphasise that even though these two methods of creating an current seem very different (one being forces applied to a wire, the other being a

changing magnetic field creating an electric field) that for either of them or any combination of the two the method of calculating the voltage by looking at the change in flux works.[1]

The fact that a changing magnetic field creates an electric field suggests that the electric and magnetic fields are more closely related than we originally thought. In fact, the rules of electromagnetism are inconsistent as we currently know them! If we kept only the rules we knew, the answers to some of our calculations would depend on how we choose to calculate them! The change that was ultimately suggested (and shown experimentally to be correct) is

> A changing *electric field* creates a *magnetic field*.

If we accept this new rule, an interesting possibility arises. If we have a magnetic field that is changing, we can create an electric field. If that electric field changes, it can create a magnetic field. We could at least imagine a situation where we start a magnetic field going, and then it creates an electric field which is itself changing, which creates a magnetic field which is changing, which ... A proper mathematical treatment shows that not only can these disturbances occur, but that these disturbances do not happen in the same place – rather they travel like the material waves we are familiar with. We call these propagating disturbances in the electric and magnetic fields *electromagnetic waves* – more colloquially referred to as *light*.

9-4-1 Harmonic electromagnetic waves

While there are many different types of electromagnetic waves, including pulse waves and spherical waves, we will devote our attention to the harmonic wave (or plane wave). The reason is a practical one, as a long distance from the source the wavefronts look flat and so the plane wave description is a good

[1] You may think that it is odd that two such disparate approaches can be summarised so easily without caring if it is "really" the magnetic field or "really" the created electric field that pushes the charges. This fact puzzled many physicists as well, until Einstein's theory of special relativity showed that the two expressions must be the same. We will not show this here, but note it as something that does seem to be crying out for an explanation.

one. For the electromagnetic wave it is the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields that oscillate:

$$\boldsymbol{E}(x,t) = E_0 \sin\left(\frac{2\pi t}{T} \pm \frac{2\pi x}{\lambda} + \phi\right) \hat{\boldsymbol{e}}$$
$$\boldsymbol{B}(x,t) = B_0 \sin\left(\frac{2\pi t}{T} \pm \frac{2\pi x}{\lambda} + \phi\right) \hat{\boldsymbol{b}}$$

The difference from our earlier expressions are the vectors $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{b}}$ at the end. Because the electric and magnetic field are vectors they must point in specific directions. The vector $\hat{\boldsymbol{e}}$ is a unit vector, and tells us the direction that the electric field is oscillating along. Similar remarks apply for the magnetic field and the unit vector $\hat{\boldsymbol{b}}$.

The directions $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{b}}$ are related. For an electromagnetic wave travelling through free space the electric and magnetic fields oscillate perpendicular to one another, and are also perpendicular to the direction that the wave travels. If you hold your thumb, index and middle fingers perpendicular to one another you can always point your thumb in the direction of the $\boldsymbol{E}$ field, your index finger in the direction of the $\boldsymbol{B}$ field, and your middle finger will point in the direction that the wave is travelling. An example of an electromagnetic wave is shown in figure 9-4.1. Because the oscillations in both the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields are perpendicular to the direction of motion an electromagnetic wave is a *transverse wave*. The *plane of polarisation* is the plane containing the direction electric field oscillates in and the direction the wave travels (i.e. the plane containing $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{k}}$ in figure 9-4.1).

Figure 9-4.1 demonstrates a couple of other points that are in our equations but we have not addressed explicitly. We notice that the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields have the same wavelength λ. They are also in phase, and so must have the same phase constant ϕ.[2] That is why we only wrote λ and ϕ in the equations, rather than placing the subscripts E and B on everything. The disturbances in the **E** and **B** fields also travel with the same speed, which tells us that the periods must be the same.

The amplitudes of the electric and magnetic fields, E_0 and B_0, are also related. The stronger the magnetic field, the more "change" in the magnetic field is going to occur as it oscillates so we would expect a bigger electric

[2]Technical aside: this statement should really be "modulo 2π".

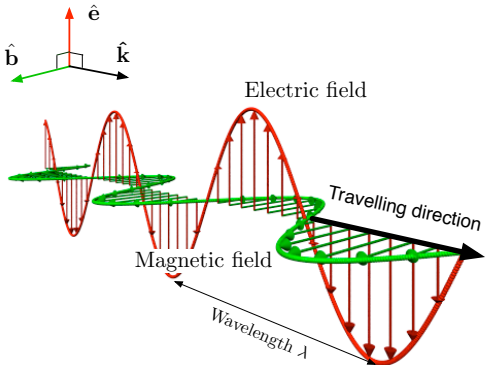


Figure 9-4.1: A diagram of an electric field at a constant time. The electric field (vertical) and magnetic field are perpendicular to each other and the direction the wave is travelling.

field. This is in fact the case. The relationship between the amplitudes of the oscillations of the electric and magnetic fields is

$$E_0 = cB_0$$

where c is the speed of light (approximately 3×10^8 m/s).

Test yourself:

If an electromagnetic wave had its electric field pointing to the right on this page, and the magnetic field pointing to the top of the page, which way would the electromagnetic wave be travelling?

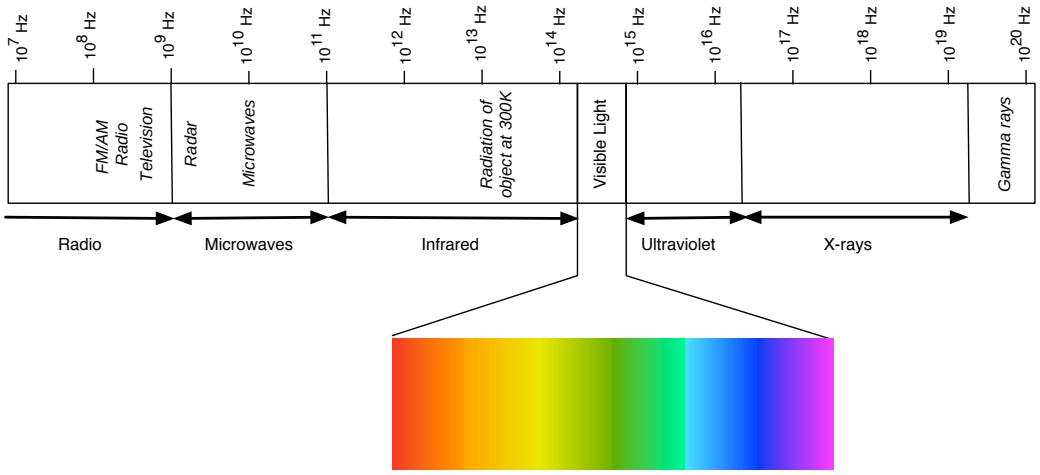
9-4-2 The electromagnetic spectrum

Our picture of the electromagnetic field being an oscillating electric field creating an oscillating magnetic field, which in turn creates an oscillating electric field, which in turn ...suggests a way of creating electromagnetic waves: taking charges (the source of electric fields) and oscillating them up and down to get the whole process started! The frequency with which the charge bounces up and down gives the frequency of the electromagnetic waves produced, similar to how the frequency of someone bouncing up and down in a pool sets the frequency of the outgoing water waves. For electromagnetic waves, like all the other waves we have dealt with, the frequency is determined by the source.

This seems a little odd, however. Most of us would have been "zapped" by a sweater we were wearing at some point in our lives, due to the build-up of charge on it. By swaying backwards and forwards we were making those charges oscillate, but we did not *seem* to suddenly create light! In Physics 7A we learned that atoms at temperatures higher than absolute zero oscillated, and these atoms are made up of electrons and protons yet most objects do not *appear* to glow in the dark. In fact, it seems if light is the oscillation of charge it should be very difficult to find darkness at all!

In fact we do give off electromagnetic radiation as we sway back and forth, and objects in dark rooms do glow. The "light" created is just not *visible light*. For visible light the charges have to be oscillating back-and-forth around 10^{14} times per second! For objects at room temperature (around 300K) the oscillating atoms give off electromagnetic light at roughly 10^{13} Hz. This light is called *infrared*, because it is right next to the red (visible) light. While people cannot see this light, some animals can and we can make cameras that can detect this light. This is how night-vision goggles work for example. For an object to give off a significant amount of visible light we must make its atoms vibrate more, and as we learned in 7A one way to do this is to increase the temperature. A wood fire, for example, burns at around 1500 K. While most of the light is let off in the infrared, enough is let off in the visible that the flames can be seen. The night sky that we see is full of electromagnetic radiation with a frequency of 3×10^{11} Hz which we cannot see directly. Just like infrared light, we have devices that can detect this light, and studying it is giving us further insights into the origin of the universe.

Different frequencies of electromagnetic waves get used for very different purposes. We are familiar with visible light already, and most of us have probably heard of infrared and ultraviolet light. At the lowest frequencies (and hence longest wavelengths because $\lambda = v_{\text{wave}}/f$) we have the radio waves. This is a broad range of frequencies lower than roughly 10^9 Hz. Most of our television and radio programmes are broadcast at this frequency. Above that we have the microwave frequencies, which are used in RADAR and microwaves. Above 10^{11} Hz we have the infrared, which are the frequencies that most objects glow most strongly in provided their temperature is between 3 K and 5000 K. The "narrow" range between 4×10^{14} Hz and 7.5×10^{14} Hz corresponds the visible spectrum. In this range different frequencies correspond to different colours of light.



At even higher frequencies we go into ultraviolet light. This region covers frequencies from 8.6×10^{14} to 3.75×10^{16} Hz and is further broken down into UVA, UVB, UVC, Far UV, and Extreme UV categories. UV light can be damaging to the skin. The risk increases with the frequency of the UV light. The sun emitted radiation in the UVA, UVB, and UVC sub-bands, however almost all of the UVB and UVC radiation from the sun is absorbed in the Earth's ozone layer in the upper atmosphere.

The use of the term X-ray varies a little. Some people take the definition of an x-ray as the manner in which the light is produced, such as an atomic transition. Others take the x-ray to simply be a frequency range like our previous definitions (this latter set tend to be astronomers talking about "x-ray telescopes"). The definition is actually fairly irrelevant, except for the purposes of communicating with others. Whichever definition we use, it is accepted that x-rays have very high frequencies (greater than 3.75×10^{16} Hz) and are energetic enough to pass through tissue. Hence we use x-ray machines to image the bones.

Gamma rays are produced in nuclear transitions, and are any photon that has frequencies higher than 10^{22} Hz.

Fire and chemistry*

We remarked that a typical wood fire has a temperature around 1500 K, which peaks in the infrared, but that some light is emitted in the red as well. For objects giving off light because of high temperatures (e.g. fire or hot metal) low temperatures mean a dull red, and as the temperature increases the frequency of the peak of the emitted light increases. The amount of

Part of Spectrum	**Typical size in vacuum**
Short wave radio	$\lambda \sim$ Building
AM/FM/TV	$\lambda \sim$ Person
Microwaves	$\lambda \sim$ Insect
Infrared	$\lambda \sim$ Flea
Visible	$\lambda \sim$ Cells
Ultraviolet	$\lambda \sim$ Molecules
X-rays	$\lambda \sim$ Atoms
γ-rays	$\lambda \sim$ Nuclei

Colour	λ_{vac} range	λ middle	Frequency
Red	620–750 nm	700 nm	4.3×10^{14} Hz
Orange	590–620 nm	600 nm	5.0×10^{14} Hz
Yellow	570–590 nm	580 nm	5.1×10^{14} Hz
Green	495–570 nm	540 nm	5.5×10^{14} Hz
Blue	450–495 nm	470 nm	6.4×10^{14} Hz
Violet	380–450 nm	400 nm	7.5×10^{14} Hz

Table 9-4.1: Labelling the different parts of the electromagnetic spectrum. Note that the frequency definitions are valid in any medium, but that when light travels into a different medium the wavelength changes. The wavelengths presented here are only correct in vacuum.

spread in frequencies also increase as the temperature goes up, so the fires always have some red in them. As the temperature goes up we go from red to orange to white, as white contains all of the colours. We do not get objects so hot they are blue – there is always some contamination from the red (and lower) part of the spectrum.

When we study chemistry we will sometime see flames of various colours as we oxidise different chemicals (e.g. in a Bunsen burner). If you have not seen this in chemistry you have probably seen it in a fireworks show – we use various chemicals in the fireworks so that we have different colours when we oxidise them. How do we reconcile this with the statement made above: as objects get hotter they get more of the visible spectrum, but still retain the "red" portion of the spectrum? The answer is that the fires we mentioned above are really thermal fires due to the random motion of charges and, more importantly, the equipartition of energy. If we have a pure substance that reacts strongly we can get chemical fires which have temperatures dictated by the energy levels available to the atoms. This effect is a quantum mechanical one that we have not studied yet, and is touched upon briefly in unit 10-1.

9-4-3 Intensity and energy of electromagnetic waves

While we have not emphasised it so far, electric and magnetic fields both contain energy. The total amount of energy depends on the values of the

fields everywhere, so it is more convenient to define the *energy density* of the fields. This is the amount of energy per unit volume the fields have, and we only need to know the value of the electric and magnetic fields *at that point*. This is similar to our motivation for introducing energy densities when we discussed fluids in Physics 7B. The energy density of the electric and magnetic fields are

$$u_E = \frac{1}{2\mu_0}\frac{E^2}{c^2}, \qquad u_B = \frac{1}{2\mu_0}B^2$$

where μ_0 is the permeability of free space introduced when we discussed Ampère's law in §9-3-3-2, $\mu_0 = 4\pi \times 10^{-7}$ N/A^2. These results are a little bit tricky to derive from what we already know, and we do not attempt a derivation here. We will just take this energy density as given.

These energy densities apply for either an electromagnetic wave *or* static electric or magnetic fields. In the case of an electromagnetic waves the energy density is the *sum* of the contributions from the electric field and magnetic field. We know that when the electric and magnetic fields hit zero then there is no energy in the electric or magnetic fields. Both the electric and magnetic fields have their peaks together, and that is where the energy density is greatest:

$$\begin{aligned} u_{\max,\mathrm{EM}} &= u_{e,max} + u_{b,max} \\ &= \frac{1}{2\mu_0}\frac{E_0^2}{c^2} + \frac{1}{2\mu_0}B_0^2, \qquad \text{(Because both E and B are maximum).} \end{aligned}$$

If we recall the result from the previous section $E_0 = cB_0$. We find that the most energy density that an electromagnetic wave has is

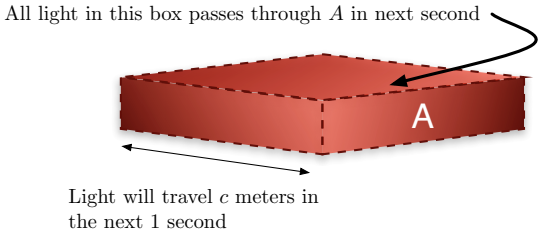
$$u_{\max,\mathrm{EM}} = \frac{1}{\mu_0}B_0^2.$$

The energy density oscillates between 0 and $u_{\max}$, we can show that the *average* energy density is half the maximum value

$$u_{\mathrm{ave,EM}} = \frac{1}{2\mu_0}B_0^2.$$

Our eyes (and other devices like film) are sensitive to how much energy comes into them per unit time, or the *power* of the light. The total power the light source is not necessarily a good measure of how bright something is – if the object is a long distance away then most of the light does not go into

our eyes (or hit the film). Instead we define *intensity* as the power of light going through a *unit* area. The more intensity the light that reaches us has, the "brighter" the light appears. To calculate the intensity, we see that all the light passing through the area A in the next one second is all the light in the box shown below:



The average energy density is $u_{\max}$, so the total energy passing out the end of this box in the next second is $u_{\max}cA$. Dividing by the area we get the intensity I:

$$I = \frac{u_{\max}cA}{A} = \frac{c}{2\mu_0}B_0^2 = \frac{1}{2\mu_0 c}E_0^2$$

9-4-4 Polarisation and polarisers

We have already discussed that an electromagnetic wave is transversely polarised, and this is fairly unambiguous because both the electric and magnetic fields oscillate perpendicular to the direction the wave travels. We also mentioned that the plane of the polarisation of the wave was in the same plane that the electric field oscillated in (i.e. the plane that has the direction of the wave and $\hat{\boldsymbol{e}}$ in figure 9-4.1.) For example, vertically polarised light would have its electric field oscillating up-and-down. On one level we can just take this as a definition, but it is useful to note why we chose to define the polarisation this way instead of using the magnetic field.

For a wave oscillating in free space it does not make much of a difference if we defined the plane of polarisation as the plane of oscillation of the electric field or the magnetic field. In fact, for an electromagnetic wave travelling in a vacuum it does not matter much what the polarisation is at all. When this wave has to interact with matter – by being absorbed, scattered, reflected or refracted – then the polarisation becomes important.

Why choose the electric field over the magnetic field for defining the polarisation? When a light ray hits an object at rest, to a good approximation the charges are more or less at rest themselves so the magnetic field from the

wave cannot exert a force on the particles. The electric field, however, can exert a force on the particles and get them to move. But then why doesn't the magnetic field become important? Well, even if the particles are moving significantly we know that they must be travelling slower than the speed of light c. Because the magnitude of the electric field $E_0 = cB_0$ for an electromagnetic wave, the magnitude of the electric force on the particles must be larger than the magnetic force on the particles! Finally we can easily figure out the direction of the electric force on particles: the force on the charge is in the direction of the field for a positive charge, or against the field for a negative charge. Compare this to the rules we would have to use to figure out the direction of the magnetic force on a charge! We choose to talk about the polarisation in terms of the electric field because the electric force is greater than the magnetic force and it is easier for us to figure out the directions.

The polarisation of a wave depends on how the wave was produced and what it interacted with. Lasers typically (but not always) produce polarised beams. Light from thermal sources (such as an incandescent lightbulb or the sun) are produced by the random vibrations of atoms, so at different times the light is polarised completely differently. We call these sources *unpolarised*. Some polarisations are preferentially reflected or absorbed by matter, so another way of polarising light is to have it interact with matter.

9-4-4-1 How a polariser works.

There are many materials which polarise light. Synthetic plastics (called *polaroids*) and natural crystals are common examples. The common feature among these materials is that they all have long linear chains of atoms which are oriented in one direction, and that the electrons in these media can travel more easily along the direction of the atomic chains. This allows the electric fields which are oriented in the direction of the atomic chains to transfer their energy to the electrons in the medium. The component of the electric field which is perpendicular to the atomic chains cannot move the electrons in that direction and is therefore transmitted. If we are given any polarisation of light we can break it into components: a component for which the polarisation of the wave is in the parallel to the chains of molecules (which will be absorbed by the electrons) and a component which is perpendicular to the chains of molecules (which will pass through as they cannot be easily absorbed). We call this perpendicular direction the *transmission axis* or *polariser axis*, as this is the direction the light that has passed through the polariser will be polarised.

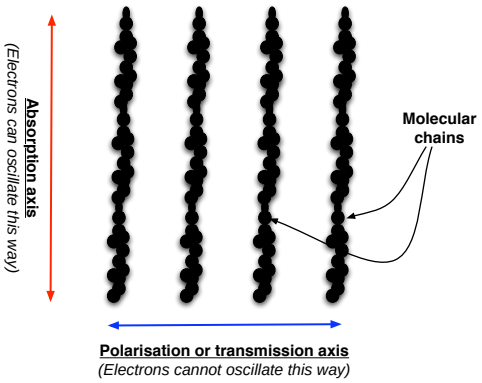


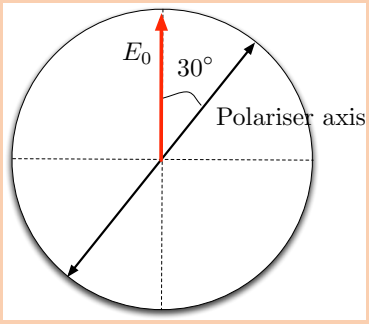
Figure 9-4.2: A diagram of a polariser. It is easy for the electrons to travel up and down the *absorption* axis along the chains, so the light's energy is easily transferred to the electrons. It is harder for the electrons to travel in the direction of the *polariser* or *transmission axis*. Any light that passes through this polariser will come out horizontally polarised.

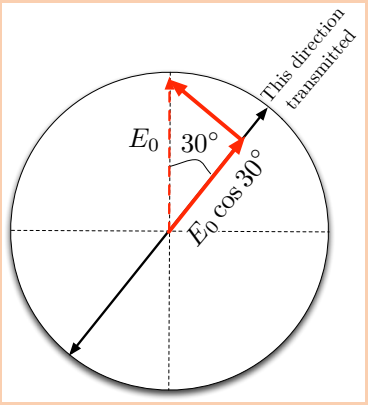
Example #1:

Vertically polarised light travelling into the page hits a polariser. The *polarising axis* is 30∘ from the vertical. What is the intensity of the light coming out of the polariser as a fraction of the incoming light? Which way is the light leaving the polariser polarised?

Solution:

We will call the amplitude of the electric field of the original wave E_0, and we know that the initial electric field is vertical. The polarisation axis and the electric field are shown on the picture to the left. On the right we break the electric field into a part that is along the polariser axis (and will be transmitted) and a part that is along the chains (and will be absorbed).





The magnitude of the electric field that makes it through the polariser has a magnitude $E_{\text{trans}} = E_0 \cos 30°$, and is oriented 30 degrees from the vertical. The intensity of the light that makes it through the polariser is

$$\begin{aligned} I &= \frac{1}{2\mu_0 c} E^2 \\ &= \frac{1}{2\mu_0 c} E_0^2 \cos^2 30° = \left(\frac{1}{2\mu_0 c} E_0^2\right)\left(\frac{3}{4}\right) \end{aligned}$$

But E_0 is just the initial intensity, so $E_0^2/(2\mu_0 c)$ is the initial intensity I_0. Therefore we have

$$I = \frac{3}{4} I_0.$$

So our final result is that the outgoing light is polarised 30° from the vertical and has 3/4 of the intensity.

In general we may be interested in how much light gets through a given polariser if we have light of a given intensity I_0 coming in. Using the same reasoning as the previous example, if the electric field is E_0 and the angle between the polarisation plane of the light and the polarisation axis of the polariser is θ then we can break the field into two parts:

$$E_{\text{blocked}} = E_0 \sin\theta, \qquad E_{\text{trans}} = E_0 \cos\theta.$$

The intensity of the light coming out is proportional to the transmitted electric field squared, so we have

$$I_{\text{out}} = I_0 \cos^2\theta.$$

This equation is sometimes referred to as Malus's law.
For *unpolarised light*, I_{out} is always half the initial intensity.

9-4-5 Do we need fields?*

When we introduced fields we introduced the idea of a direct method of calculating forces and a field method for calculating forces. At the time it seemed like the field was little more than a book keeping device, and that if were willing we could use the direct method between all the particles that we were interested in. Even when we discussed magnetic fields and did *not* give a direct method because the field method was far easier, it seemed like it was possible to construct a direct method with enough patience. Nothing we discussed earlier seemed to *require* fields, fields just made the job easier. One might be tempted to ask if fields are convenient mathematical constructions.

To show why fields (or something like them) have to exist, consider moving a charge around for a short time. Moving the charge around requires that you give the charge some energy, and that charge radiates away that energy as electromagnetic radiation. Sometime later the wave hits some other charges, and starts them oscillating. At all stages energy is conserved.

What would happen if we did not know about fields? In that case we would see the charges that we originally dumped energy into slow down and eventually stop. It seems like we lost energy, but we might not worry about that – after all it could have lost that energy to collisions with other atoms. But then sometime later the other set of charges would start to oscillate all by themselves! It would seem like we lost some energy for a short time, and then got it back again later. This would lead us to three possibilities:

1. Maybe the conservation of energy is not valid (**wrong**).
2. Maybe the energy went into an energy system (such as a field) that we didn't count, and everything is really okay.

We have found energy conservation immensely useful so we would hate to abandon it unless we absolutely had to. Obviously the choice that we are advocating is the second one, but some may still object. There is a third possibility that allows us to not have fields but still conserve energy:

3. Maybe the energy of a system does not only depend on it is doing right now, but also what happened to it in the past. That is, maybe energy is not a state function.

This is a very clever attempt to get around having to accept fields, but unfortunately it fails. The *whole point* of energy conservation is that the

energy is a state function. If statement 3 were true we would have to know the entire history of the universe (in principle) to figure out what the energy should be now. The point of energy conservation is you only have to know about the initial (or final) time and do not need to do the hard work of figuring out exactly what happened in between.

So in order to save the principle of the energy conservation in a *useful* manner, we are forced into accepting that there is an additional energy system – the field. The price we pay is that when we choose our initial and final times we should pay at least a fleeting thought to how the energy in the electric and magnetic fields have changed.

9-4-6 Summary

1. We learned in §9-3-4 that a changing magnetic field causes a changing electric field. Now we learn that a changing electric field causes a magnetic field.

2. The field lines from an electric field created by a magnetic field form a loop. The idea that electric field lines only beginning or end on charges is still valid.

3. The idea of voltage is not defined in the case of a magnetic field creating an electric field. This is because the amount of energy a particle has does not only depend on where it is, but on its past history.

4. The electric and magnetic fields are very closely related, and sometimes referred to as the *electromagnetic field*.

5. An electromagnetic wave describes light. The different frequencies of electromagnetic waves determine the colour of the light (if it is visible). Unlike the other types of waves we have already discussed an electromagnetic waves does not require a medium, which is why we can see through empty space.

6. Electromagnetic waves are transversely polarised. Because there are many directions perpendicular to the direction of travel we still ask about the polarisation (direction) of an electromagnetic wave. The definition of the polarisation direction is the direction that the **E** field is parallel (or anti-parallel) to.

7. Intensity as energy per unit area, and the brighter the light the higher the intensity.

8. Become familiar with how polarisers work, and to be able to determine the brightness and polarisation of light after it passes through a polariser.

9. Fields are not just mathematical constructions, but are required to conserve energy and momentum.

Exercises

1. We learned in §9-4-3 that the average energy density in an electromagnetic wave was $u_{\text{ave, EM}} = \frac{B_0^2}{2\mu_0}$, where B_0 is the amplitude of the magnetic field. How would you write $u_{\text{ave,EM}}$ in terms of the electric field's amplitude E_0?

2. If I want horizontally polarised light to pass through a polariser, but have 20% of its original intensity, what possible polarisation angles θ may I use? What is the polarisation of the wave that leaves?

3. I start with horizontally polarised light, as in the previous case. This time, I have *two* polarisers (although I can choose to only use one). I want the light to come out after passing through the polariser(s) to have 20% of its original intensity *and* be horizontally polarised. Is this possible? If so, describe how to set the polariser(s) up.

Unit 10

Quantum

Unit 10:

10-1: Quantum mechanics

10-1-1 Introduction

Let us start with something familiar. Take a cup. Try filling it with water. Now, we know that we cannot put any more than one cup of water in there, right? So if someone asked you how much water could go in the cup you would probably respond with an answer like "I can put any amount less than a cup in here."

But we also know that water is H_2O molecules. A standard cup is 250 mℓ, which is roughly 8.4×10^{24} molecules of water. The number is not important, but you should know how to calculate it (the calculation at the end of this section). But water molecules cannot be split in half and still called water! So the amount of water you can fit into a cup cannot take any value: you must have 0 molecules of water, 1 molecule of water, ... or 8.4×10^{24} molecules of water in the cup. We say that the amount of water in the cup is *quantised* as we can only fit certain allowed *quantities* of water into the cup.

In quantum mechanics, many of the quantities we deal with such as energy and angular momentum can only take certain values as well. The word "quantum" is derived from the Latin word *quantus*, the same root word as quantity. The name quantum mechanics comes about to remind us that many of the things that we deal with only come in certain quantities. So far we have considered objects interacting by exchanging energy, momentum or angular momentum with one another assuming that we could transfer any amount of energy, momentum or angular momentum we wished. We will focus on quantising the *energy* and see what the consequences are.

With the example of quantised water molecules, the picture we have of molecules makes it easy to visualise why water is quantised. We also know that any time we have water, the "quantum" (i.e. smallest amount) will

always be one molecule. What about something abstract like energy? What picture should we have in mind, and is the "quantum" of energy always the same? The answer to this final question is **no**; the *allowed energies* of a system depend on the situation. For example, an atom has different allowed energies that a mass on a spring. [1] One of the mathematically more difficult parts of quantum mechanics is finding the allowed energies of a given system.

Example #1:

How many molecules are there in a cup (250 mℓ) of water?

Solution:

This solution uses the definition of moles and atomic weights as covered in Physics 7A.

We know that the density of water is 1 g per cubic centimeter, or 1 g/mℓ. A cup of water is 250 mℓ, so the mass of water is

$$m_{\text{cup of water}} = \rho V = \left(\frac{1\,\text{g}}{\text{mL}}\right)(250\,\text{m}\ell) = 250\,\text{g}$$

The molecular weight of a water molecule is given roughly by the fact that oxygen has 16 nucleons (8 protons and 8 neutrons), and hydrogen has 1 nucleon. The total molecular weight is 16 + 1 + 1 = 18. This means that each mole of water molecules has a mass of 18 grams. The number of moles in a cup of water is

$$250\,\text{g} = \frac{18\,\text{g}}{\text{mole}} \times n_{\text{moles}}$$

Therefore the number of moles is 13.9. Each mole has an Avogadro's number of particles by definition. So the total number of water molecules is

$$\begin{aligned} N_{\text{molecules}} = N_A \times n_{\text{moles}} &= \left(6.02 \times 10^{23}\,\frac{\text{particles}}{\text{mole}}\right) \times (13.9\,\text{moles}) \\ &= 8.4 \times 10^{24}\,\text{particles} \end{aligned}$$

So a cup of water has 8.4×10^{24} molecules of water in it.

[1] Technical aside: the angular momentum a particle is allowed to have depends on the type of particle, but not on the potential that particle is placed in.

In the quantum unit our plan is to *tell* you the allowed energies for three systems:

- A simple harmonic oscillator (e.g. mass-spring, atomic bonds)
- The hydrogen atom
- A particle confined to a box

We study these three systems to get an idea of what the *consequences* of only being allowed to transfer specific amounts of energy between systems are. After this we deal with how energy gets quantised in the first place, and some of the strange nature of very small objects. The reason we take this approach is that in science we are trying to find the *simplest set of assumptions* to describe the world. Taking energy levels for granted helps us explain phenomena relevant to biology (e.g. photosynthesis) and chemistry. The later sections §10-1-3 discusses *why* energy is quantised in the first place is a lot more complicated, but leads to more powerful predictions. This is similar to the approach we took in Physics 7A – we started with the three-phase model and then later developed the ideas of atoms and modes to explain (at least some of) the reasons why we had a three-phase model.

10-1-2 Quantised energies

Until this unit, our model of energy allowed a particle to have any energy value. In the quantum mechanics model, this is still true of particles moving freely through space, but the energy of a *confined* particle is *quantised* – meaning only certain values of energy are allowed, as discussed in the introduction. Like other models developed in this volume, understanding a few key ideas about quantised energy levels will enable us to make sense of a variety of phenomena, from the emission spectrum of the hydrogen atom to the unfreezing of modes in vibrational atoms.

When we describe the energy of a particle as quantised, we mean that only certain values of energy are allowed. Perhaps a particle can only have 1 Joule, 4 Joules, 9 Joules, or 16 Joules of energy. In this case, whenever we measure the particle's energy, we will find one of those values. If the particle is measured to have 4 Joules of energy, we also know how much energy the particle might gain or lose. The particle could only gain the exact right amount of energy to get to a higher energy level–in this case, it could gain 5 Joules (to reach the 9 J level) or 12 Joules (to reach the 16 J level). No

other amount of energy could be added, other than 5 Joules or 12 Joules.[2] Similarly, if the particle were to lose energy, the only amount it could lose is 3 Joules, leaving it with 1 J. The energy state with the least amount of energy is called the *ground state*.

How does a particle change its energy level? We learned in Physics 7A that if we consider all the energy systems involved in a process then the energy is conserved. If a particle goes to a "higher" energy level (i.e. gains energy) that energy must come from somewhere.[3] Likewise, if a particle goes to a "lower" energy level, that energy must be transferred into another system. The most common systems for small systems of particles to transfer energy to or from are vibrations of molecules or light.

We know that energy is always conserved if we include all energy systems involved in the process. So where did this energy come from, or where did it go? The mechanism for conserving energy can be anything from heat (often the case in vibrating molecules) to light (the usual mechanism in transitions of electrons). We will talk about the energy of light in more detail in §10-1-2-2.

Every system has a collection of allowed energy levels is called the *energy spectrum*. These energy levels tell us the *total energy* the particle is allowed, which is usually the sum of the kinetic and potential energies. The allowed energies depend on the potential energy of the system ($PE(r)$) and the mass. In effect, every potential energy forms a "container" of sorts that confines the particle to a specific area. Each unique container has its own set of energy levels. We will develop this idea with more examples below in §10-1-2-1.

As discussed earlier in this volume and in previous quarters, the potential energy of a particle is not important. It is the *changes* in potential energy that matter. In quantum mechanics, this is still the case. Though we can set the zero of potential energy at any place, we make conventions that make the interpretation as simple as possible.

[2] In actuality, there are usually an infinite number of allowed states, not just the four.

[3] Because the energy level is the total energy, we cannot have energy transferred from PE to KE (for example). Changing the energy level means interacting with another system.

Test yourself:

In the example mentioned in the text where the particle can have 1 J, 4 J, 9 J or 16 J of energy, what is the energy spectrum? What is an example of an energy level?
(Note: there is no physics involved here, just a check that you have the definitions under control.)

Quantisation of light: photons

Just as we can think of ordinary matter being quantised (as illustrated by the example with water), we also find that light comes in indivisible "packets" that we refer to as *photons*. Unlike atoms, we do **not** model these photons as being made up of smaller particles. We will discuss photons in more detail in §10-1-2-2 , but at the moment we only need a couple of facts about photons.

1. The energy of a photon depends on its frequency f:

   $$E_{\text{photon}} = hf,$$

   where $h = 6.636 \times 10^{-34}$ J s is *Planck's constant*, a universal constant. We know from §9-4-2 that the frequency is related to the type of light we are dealing with. In particular, different colours correspond to different frequencies and therefore have different energies.

2. A photon is the smallest amount of light energy we can get (at a particular frequency).

A useful analogy is to again consider different elements: to a good approximation we can think of mass as being quantised, with the mass of individual atoms being the "indivisible" unit of mass (after all, it is hard to break up atoms!). However different elements have different masses. Each photon is an indivisible unit of light energy, but different frequencies have a different "fundamental amount" of energy. The light that we deal with everyday is typically made up of many photons, so we don't notice the individual photon nature of light any more than we notice the individual atoms in the materials we use everyday.

10-1-2-1 The potential determines the energy spectrum

Gaining a better understanding of three important energy spectra will help us learn about a large variety of phenomena. The first spectrum we will consider is called the *infinite well*. In this system, a particle is trapped in such a way that it can only be between two points, no matter how much energy it has. We contrast this with a *harmonic oscillator* (like a mass on a spring), where having more energy enables a particle to travel through more distance. The third system we will explore is a system with a single electron bound to a nucleus.

An example of an energy spectum for each of these three energy systems is shown in figure 10-1.1. We will examine each in more detail. By convention, the vertical axis represents energy. The horizontal has no meaning. In looking at these spectra, focus on how the spacing between consecutive energy levels is different in the three situations. Note that the levels are evenly spaced for the oscillator (center), closer together at low energies for the infinite well (left) and closer together at higher energies for the hydrogen atom (right).

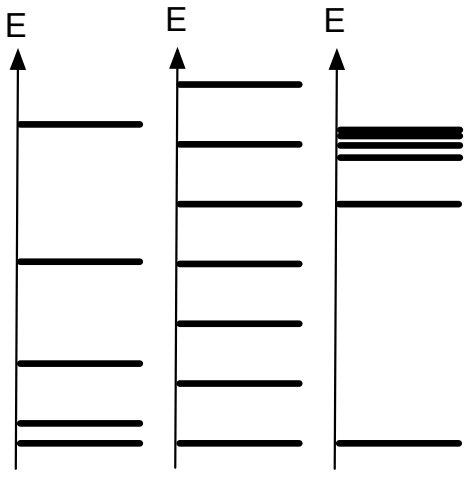
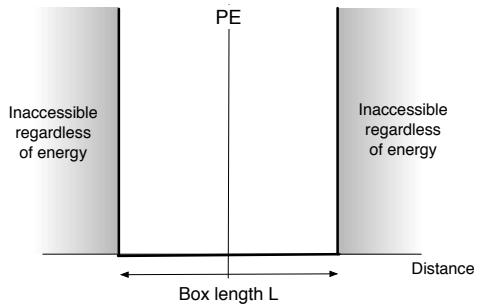


Figure 10-1.1: On the left is the energy spectrum of the infinite well, in the center the spectrum of a harmonic oscillator, and on the right the spectrum of a electron in a hydrogen atom.

The infinite well

The infinite well is a system where a particle is trapped in a box of fixed size, but is completely free inside the box. To keep the particle trapped in the same region *regardless* of the amount of energy it has, we require that the potential energy is infinite outside some region. We will choose to take the potential energy to be zero inside the box.



The infinite well seems to be the least useful of the situations we will study. Very few physical situations are similar to the infinite well. We introduce this system because it has the simplest potential available. If a particle is inside the box then it has no potential energy. If the particle is anywhere else, it has infinite potential energy. Because our particles can only have finite energy this ensure the particle stays in the box. Also, since there is zero potential energy inside the box, the total energy of the particle is equivalent to the kinetic energy of the particle. If the particle gains total energy, we know it must have gained kinetic energy.

In §10-1-3-5 we will develop the formula for the energy levels of a particle trapped in a square well. For now, we will make sense of the equation without worrying about its derivation. Suppose that the length of the box is L, and the particle trapped in this potential has a mass m. Then the allowed energies are

$$E_n = n^2 \frac{h^2}{8mL^2}$$

for any positive integer n, and h is Planck's constant again. Because $n \geq 1$ we see that the minimum amount of energy the particle can have is $E_1 = \frac{h^2}{8mL^2}$.[4] Because the potential energy is zero inside the box, we see the particle *always* has some kinetic energy, and is always vibrating. For a quantum particle in a box it is *impossible* to make it sit at rest.

[4]Strictly speaking, we could always redefine where we put the zero of potential energy so $E_1 = 0$. However, the critical part is the kinetic energy would be non-zero, so the particle would still be vibrating.

Example #2:

Suppose you have a particle in the ground state of the infinite square well potential. You also have a device allowing you to add energy to the particle. You start by trying to add an infinitesimally small amount of energy, and nothing happens. If you slowly increase the amount of energy you try to add to the particle, what is the smallest amount of energy you can successfully transfer?

Solution:

We know that the particle's energy is quantized, and that the only allowed levels are $E_n = n^2 \frac{h^2}{8mL^2}$. If the particle begins in the ground state, then $n_{\text{initial}} = 1$. Adding very small amounts of energy will not have any impact on the particle's energy, because the particle cannot gain energy unless it can transition to the next higher energy level, when $n_{\text{final}} = 2$. The energy we must add corresponds to the difference in energy levels:

$$\begin{aligned} \Delta E &= E_{\text{final}} - E_{\text{initial}} \\ &= E_2 - E_1 \\ &= (4-1)\frac{h^2}{8mL^2} = 3\frac{h^2}{8mL^2} \end{aligned}$$

The *only* way the particle will make the transition from the ground state to the first excited state is if something transfers $3\frac{h^2}{8mL^2}$ into it, giving the particle just the right amount of energy to make the transition.

For the infinite well potential, the energy levels are proportional to n^2. This means that it is easier to transition between the lower energy levels and harder to transition between higher energy levels, where the energy gap is larger. Also, the energy gap between consecutive levels is smaller if L is bigger, corresponding to a wider potential.

Example #3:

The protons and neutrons of an atom are, to a good approximation, confined to the nucleus. We will model the nucleus as an inescapable box of size 10^{-15} m (typical for atomic nuclei). Give an estimate of how much energy we would need to move a proton in Helium up to the next energy level.

Solution:

From our previous example we know that the amount of energy to go from

the ground state to the second state is

$$\Delta E_{1\to 2} = 3\frac{h^2}{8mL^2}$$

Putting in numbers we get

$$\Delta E_{1\to 2} = 3\frac{(6.626\times 10^{-34}\,\text{J s})^2}{(8)(1.67\times 10^{-27}\,\text{kg})(10^{-15}\,\text{m})^2} = 9.9\times 10^{-11}\,\text{J} \approx 6\times 10^{8}\,\text{eV}$$

How did we know the proton would start in the ground state? We didn't! This is a rough estimate only, but it gives us some idea of the amount of energy involved. To make a meaningful comparison, the amount of energy it takes to break a chemical bond has a typical magnitude of 1 eV.

Simple harmonic oscillator

The next potential we will consider is the harmonic oscillator. A microscopic particle that is constrained by a spring-like potential (for instance, the atom-atom potential) will also have quantised energy levels. Although there are no actual springs, we still introduce the idea of a spring constant in describing the potential energy:

$$\text{PE} = \frac{1}{2}k_{\text{spring}}x^2$$

As we learnt in Physics 7B, a mass on a spring with spring constant k_{spring} would oscillate with a frequency f given by

$$f = \frac{1}{2\pi}\sqrt{\frac{k_{\text{spring}}}{m}}.$$

Recall that this frequency is independent of the amplitude of the oscillation.

For this potential, the energy levels are equally spaced, and the spacing is related to the frequency of the oscillation.

$$\begin{aligned} E_{\text{n}} &= KE + PE_{\text{spring}} \\ &= \left(n-\frac{1}{2}\right)hf; \qquad n = 1,2,3,4,5,... \\ &= \left(n-\frac{1}{2}\right)\frac{h}{2\pi}\sqrt{\frac{k_{\text{spring}}}{m}} \end{aligned}$$

As before, n is a positive integer, and h is Planck's constant. Because n is a positive integer we see that it is not possible for a particle to be at rest in

a mass-spring system. The potential energy in the atom-atom potential is similar to the potential energy for mass-spring systems – our atoms are oscillating even at 0 K! This energy is not particularly useful because we cannot make it *do* anything; their are no energy levels with less energy available to the mass so we cannot transfer this energy to another system.

Test yourself:

A friend of yours points out that a mass-spring in the lab would also have quantised energies due to this formula. Because the energy depends on amplitude, this would mean that you are only allowed certain amplitudes. Yet in the lab, it seems that you can set the amplitude to an value you want. How do you resolve this discrepancy? (Hint: you will want to consider the numerical value of h).

The quantisation of energy also helps us understand the freezing of vibrational modes that we learnt about in Physics 7A. Let us consider a diatomic molecule that would vibrate at a frequency f. In Physics 7A we learned the "typical" amount of thermal energy available per mode is $\frac{1}{2}k_BT$, where k_B is Boltzmann's constant, 1.38×10^{-23} J/K, and T is the temperature in Kelvin. To get an atom to *vibrate* we need to activate 2 modes – one potential and one kinetic. Therefore to start a vibration the amount of energy we need is around k_BT. The amount of energy that would need to be transferred to an atom to get it to the next state is $\Delta E = hf$, as we will show in Section 10-1, ex. #4. If the thermal energy k_BT available is *less* than hf, the diatomic molecule does not have enough energy to make it go up an energy level and it is stuck. We say the vibrational modes are *frozen out* because we cannot put any energy into that mode. When the amount of thermal energy available per mode ($\frac{1}{2}k_BT$) available per mode becomes higher than the energy gap hf between energy levels, then we can transfer energy into the vibrational energy of the atoms, and we say the vibrational mode has been *activiated*. The thermal energy per mode is controlled by the value of temperature T.

Test yourself:

Oxygen gas O_2 has a vibrational frequency of 5×10^{13} Hz. At roughly what temperature does the vibrational mode become activated?

It is not surprising that even among systems that can be modeled as simple harmonic oscillators there are some variations. Classically, we know that how

stiff or loose a spring is will affect the motion. Quantum mechanically, we find the same thing.

Example #4:

Compare the energy spectra of a vibrating molecule with a strong bond to a weak bond, assuming the masses in each case are the same.

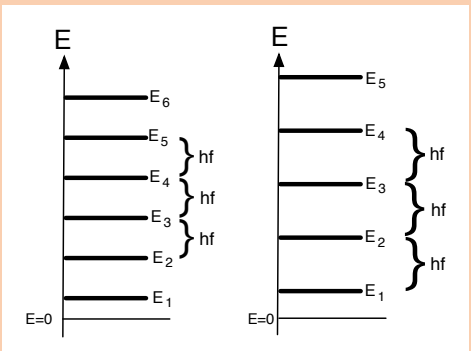
Solution:

As the masses are the same, the strength of the bond is the only parameter affecting the the energy spectra. The bond that is stronger has a bigger k_{spring}, which results in a higher frequency. The energy of the ground state is $E_{\text{ground}} = \frac{1}{2}hf$, so the molecule with higher frequency has a higher ground state energy. We will keep this in mind when we draw the energy levels.

The energy levels of a harmonic oscillator are evenly spaced, meaning that the energy required to transition from one level to the next is the same regardless of what level you're on. We can find this spacing by subtracting the n^{th} energy from the $(n+1)^{\text{th}}$ energy:

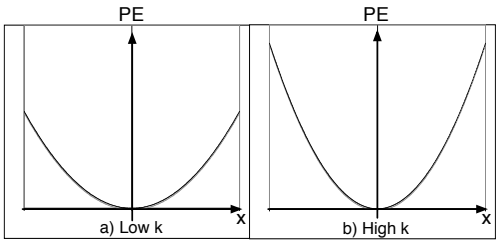
$$\begin{aligned}\Delta E &= \left((n+1) - \frac{1}{2}\right) hf - \left(n - \frac{1}{2}\right) hf \\ &= \left(n + \frac{1}{2}\right) hf - \left(n - \frac{1}{2}\right) hf \\ &= hf\end{aligned}$$

The spacing of the levels is proportional to the frequency, so the molecule with a higher k_{spring} has energy levels spaced further apart. We can use the information about ground state energies and energy spacing to show the energy levels:



The diagram above shows the energy levels of the weaker bond (left) compared to the stronger bond (right). Recall that E_1 is called the ground state energy, and note that the molecule with higher frequency vibrations has a higher ground state energy. The energy levels for the higher frequency vibrations are also spaced further apart.

The example above compares the energy levels for low frequency oscillations to those for high frequency oscillations. Because we know the potential energy determines the energy spacing, it is worth examining the differences in the potential energies for the two systems. From earlier courses, we know the potential energy in a mass-spring system is $PE_{\text{spring}} = \frac{1}{2}k_{\text{spring}}x^2$. Thus, having a higher k_{spring} also changes shape in that the stronger spring has a steeper potential (see below).



Notice that the flatter potential has energy levels that are more closely spaced. This is similar to the earlier finding that the wider infinite well has closer energies.

Energy is quantised for a diatomic molecule, and diatomic molecules can only emit or absorb quantized amounts of energy (in the form of a photon or heat). Remember that we are treating the atoms themselves as our microscopic particle; additionally there are electrons in the atoms themselves that also have quantised energy levels. Make sure you understand that the quantized energy levels for the electrons in the atoms are separate and distinct from the quantised molecules for the atoms undergoing simple harmonic motion.

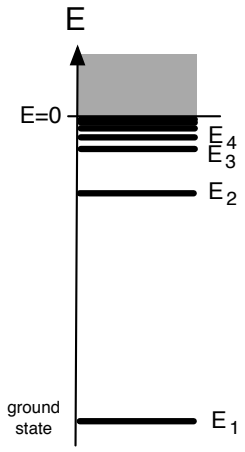


Figure 10-1.2: The spectrum for hydrogen. There are different bound energies (shown as the thick vertical lines) that get closer together as the energy gets higher. Once the electron is free from the atom ($E > 0$) the energy can be any positive value, represented by the shaded block.

Single electron in an atom

The final potential we will discuss here is an electron bound to an orbit around a nucleus. The total energy levels are quantized in terms of n as such:

$$\begin{aligned} E_{\text{n, total}} &= KE + PE_{\text{electric}} \\ &= -\frac{1}{n^2}\left(\frac{2\pi k_{\text{e}} Z e^2}{h}\right)^2 \frac{m_{\text{e}}}{2} = \frac{E_1}{n^2}; \qquad n = 1, 2, 3, 4, 5, \ldots \end{aligned} \tag{10-1.1}$$

where n is any non-zero integer; k_{e} is the electrostatic constant ($9.0\times10^9 \frac{\text{Nm}^2}{\text{C}^2}$); Z is the number of protons in the nucleus; e is the charge of the electron; h is Planck's constant; and m_{e} is the mass of the electron. The energy levels E_n can be rewritten in terms of the ground state energy E_1 as $E_n = \frac{E_1}{n^2}$. Writing out the first few energy levels explicitly

$$E_{\text{n}} = E_1, \frac{1}{4}E_1, \frac{1}{9}E_1, ...$$

Plugging in the values of the constants in equation 10-1.1, we find that $E_1 = (-2.18 \times 10^{-18}\,\text{J})Z^2 = (-13.6\,\text{eV})Z^2$. The most relevant example is the hydrogen atom ($Z = 1$), as this is the only atom that *typically* has only one electron. The energy levels are represented in figure 10-1.2

Note that all the allowed energy levels for the electron in a hydrogen atom are negative. This comes about because the potential energy of an electron bound to a hydrogen atom is chosen to be negative (revisit §9-2-4 if you have forgotten). The ground state energy level ($n = 1$) is the most bound state. Adding (allowed) energy to the electron will increase n, making its total energy less negative, or even zero (as n approaches ∞). At this point, the electron will be unbound and free, and then will be allowed to have any (positive) value of energy. Because we cannot draw a line for *every* positive energy, we have simply added a shaded region to the spectrum. Also note that the energy levels get closer and closer together for larger n.

There are many transitions that occur that allow the electron to remain bound to the nucleus. Each allowed transition requires a certain amount of energy input (or releases a certain specific amount of energy, if the electron loses energy). These interactions typically occur by the atom absorbing (or emitting) individual photons. Because only certain energies can be absorbed (emitted), and the photon's energy depends on frequency, only certain frequencies of light will be absorbed (emitted).

Example #5:

Light including the infrared, visible, and ultraviolet hits a bunch of hydrogen atoms at nearly 0K. The light is detected after hitting the atoms. a) Describe the light. Considering only transitions that allow the electron to remain bound the the nucleus, b) determine the longest possible wavelength absorbed, c) determine the shortest possible wavelength that could be absorbed, and d) determine if either of the photons in (b) or (c) is in the visible range.

Solution:

a) The light incident on the hydrogen atoms includes a full range of frequencies, and thus a full range of energies. When the light hits the hydrogen atoms, some of the photons with exactly the right energy will excite the electrons into higher energy levels. The other photons will pass through unimpeded. Thus, the light reaching the detector will no longer contain the full range of frequencies, but will have "narrow bands" missing corresponding to transition energies in the hydrogen atom, and most of the light at high frequencies absorbed by ionizing atoms. The dark bands could be observed by passing the light through a prism to separate the light by frequency.

b) Longer wavelengths of light have lower frequencies. The longest wavelength absorbed will correspond to the lowest frequencies, and thus the lowest energy transition. We start with atoms that are very, very cold, so we can assume that all of the electrons are initial in the ground state. The lowest energy transition is from the ground state (n=1) to the n=2 level. For the electron to make this transition, it must absorb the energy of a photon with the correct amount of energy.

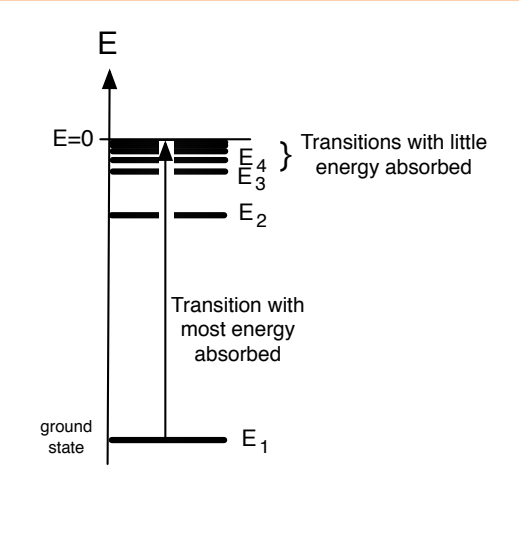
$$\begin{aligned}
\Delta E_{\text{photon}} + \Delta E_{\text{electron}} &= 0 \\
E_{\text{photon, final}} - E_{\text{photon, initial}} &= -(E_{\text{electron, final}} - E_{\text{electron, initial}}) \\
0 - E_{\text{photon, initial}} &= -(\frac{-13.6\,\text{eV}}{2^2} - (\frac{-13.6\,\text{eV}}{1^2})) \\
E_{\text{photon, initial}} &= 10.2\,\text{eV}
\end{aligned}$$

The question asks us to determine the wavelength of the absorbed light, so we must determine the wavelength of light corresponding to 10.2 eV of energy. The light is in vacuum (we are dealing with subatomic scales here, so "medium" does not make sense) so $v_{\text{light}} = c$.

$$\begin{aligned}
E_{\text{photon, initial}} &= 10.2\,\text{eV} \\
hf &= 10.2\,\text{eV} \\
\frac{hc}{\lambda} &= (10.2\,\text{eV}) \\
\frac{(6.626 \times 10^{-34}\,\text{Js})(3 \times 10^{8}\,\text{m/s})}{\lambda} &= (10.2\,\text{eV})\frac{1.6 \times 10^{-19}\,\text{J}}{1\,\text{eV}}
\end{aligned}$$

The longest absorved wavelength is 1.22×10^{-7} m or 122 nm.

The Hydrogen atoms in the problem were at nearly absolute zero, so the electrons were all in the ground state. This is not always the case. We could have an electron start in the n=2 state and transition to the n=3 state. An electron undergoing this n=2 to n=3 transition would have a longer wavelength than the 122 nm found above (calculate it and check!). In fact, the energy levels in a hydrogen atom are very closely packed near zero. There are infinitely many energies available just below zero energy, at large n, so the gap between energies can be infinitely small. Note that realistically it is unlikely that there are lots of electrons with initial states with very high n, because these electrons would be easily dissociated from their nuclei. However, for excited electrons in atoms, there is no such thing as the smallest possible wavelength absorbed.



c) Shorter wavelengths of light correspond to higher frequencies and higher energies. The highest energy transition available is from $n = 1$ to very large n, just before the electron is freed from the atom. The initial state is the ground state, with energy $-13.6\,\text{eV}$, and the final state approaches 0 eV. The electron gains 13.6 eV of energy so the photon must lose the same amount of energy. Mathematically,

$$\begin{aligned}
\Delta E_{\text{photon}} + \Delta E_{\text{electron}} &= 0 \\
E_{\text{photon, final}} - E_{\text{photon, initial}} &= -(E_{\text{electron, final}} - E_{\text{electron, initial}}) \\
0 - E_{\text{photon, initial}} &= -(0\,\text{eV} - (-13.6\,\text{eV})) \\
E_{\text{photon, initial}} &= 13.6\,\text{eV}
\end{aligned}$$

As in part (b), we must determine the wavelength of light corresponding to 13.6 eV of energy. Recalling $v_{\text{light}} = c$.

$$\begin{aligned}
E_{\text{photon, initial}} &= 13.6\,\text{eV} \\
hf &= 13.6\,\text{eV} \\
\frac{hc}{\lambda} &= (13.6\,\text{eV}) \\
\frac{(6.626 \times 10^{-34}\,\text{Js})(3 \times 10^{8}\,\text{m/s})}{\lambda} &= (13.6\,\text{eV})\frac{1.6 \times 10^{-19}\,\text{J}}{1\,\text{eV}}
\end{aligned}$$

The shortest absorbed wavelength is 9.14×10^{-8} m or 91.4 nm.

d) There is no longest wavelength (part b), so the wavelength could get arbitrarily large and certainly longer than any in the visible range. The

smallest wavelengths human eyes can see are around 400nm, so the shortest absorbed wavelength is also outside the visible range. If we convert the wavelengths back to frequencies, we find the lower frequency (c) is 2.46×10^{15} Hz and the higher (d) is 3.28×10^{15} Hz. Referring to §9-4-2 we find that both of these photons are in the ultraviolet range.

The previous example explores the *absorption spectrum* of hydrogen. We could also explore the *emission spectrum*. If we heat up a tube of hydrogen case, many of the electrons are excited out of their ground states and into higher energy states. As these electrons fall to lower energy levels, photons are emitted to conserve energy. The emission spectrum of hydrogen can be directly calculated from the energy levels. The emission spectrum for other elements are more complicated to calculate because other elements have multiple electrons that all interact with one another, in addition to the interaction with the nucleus. Because each atom has a different energy spectrum, the emission spectrum of each element is unique. The uniqueness of the spectra can be exploited in spectroscopy by identifying unknown atoms and molecules by the energies of photons that their elements emit or absorb.

Burning samples of chemicals is another way to excite electrons. We might expect to see some correlation between the emission spectrum and the color of chemical fires. In practice, we do indeed see this similarity. The color of burning elements is due to the emitted photons. To take a specific example, burning sodium produces a bright yellow flame. We can understand this by studying the emission spectrum of sodium, which includes many photons but only two in the visible range (at least at low pressures). Both of the visible photons have wavelengths of about 590 nm which corresponds to yellow light. The color of the flame is yellow because of these yellow spectral lines. You may be familiar with the color of excited sodium from sodium vapor street lamps, which also emit yellow light!

10-1-2-2 Photons

Previously, we devoted all of 9-4 to understanding the phenomena of light waves. To review briefly, we found that light behaves just like other waves in a variety of circumstances, such as when sent through small thin slits as in §8-2-6. Prior to the two-slit experiments, physicists had been uncertain about the nature of light. Prominent physicists, including Sir Isaac Newton, strongly believed that light was more like a particle than a wave. The two-

slit interference patterns of light could be understood so well with the wave model that for awhile the subject was laid to rest.

In the early 20th century, several circumstances involving light brought the wave model back into consideration. Eventually, enough evidence accumulated to conclude that light behaves in ways that can be explained by a particle model, but cannot be explained by a wave model. Presently, we must hold in our minds both the wave model of light and the particle model of light. In some circumstances, the behavior follows the wave model, but in other circumstances, it follows the particle model.

What is the particle model?

Just as energy is quantised, so is light. The individual quanta of light are called "photons." The photons can be thought of as packets, bundles, or particles of light. From our earlier discussion we know that the energy of a photon is proportional to the frequency f of the light $E_{\text{photon}} = hf$, where h is Planck's constant.

To consider the implications of the particle model, it is helpful to think about monochromatic light, meaning light with all the same frequency, like that produced by a laser. We consider two properties of the light–it's intensity (i.e. brightness) and the amount of energy the light is able to transfer into another system, like an electron orbiting a nucleus.

First, compare two beams of light with equal brightness but different frequencies. From our relationship $E_{\text{photon}} = hf$ we see the beam with the higher frequency has higher energy photons. Thus, the high frequency beam is capable of transferring larger amounts energy into another system. Because the intensity refers to the total energy in the beam, the intensity being the same tells us that the beam with the higher frequency has *fewer* photons. In the wave model, we would have said that since both beams have the same intensity, they must have the same amplitude. The energy in a wave is related to the amplitude, so both light beams must have equal ability to transfer energy. Clearly, the two models lead to different hypotheses.

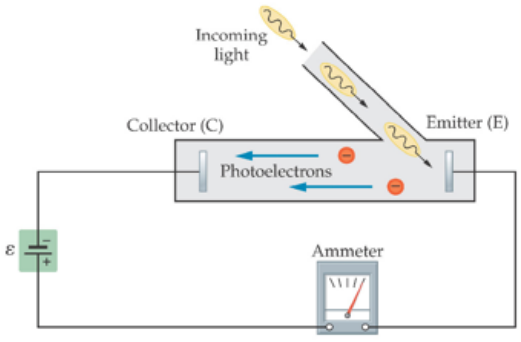
Next, consider the action of increasing the beam's brightness. In the particle model, we say that we have added more photons to the beam, but that each particular photon still only carries a certain amount of energy. Using the particle model, we conclude that the brightness of the beam does not

influence how much energy any particular photon can transfer to another system, though with more intensity we have more photons available to make the transfer. In the wave model, we would conclude that the greater intensity wave has the ability to transfer larger amounts of energy into another system. Again, the models make different predictions.

At this point, we have two different models for light. We know that the wave model is quite able to predict the behavior of light in two slit interference, where the particle model can not. So why learn about the particle model? The photoelectric effect is an experiment that provides strong experimental evidence of the particle model of light. In fact, it was the photoelectric effect that first led Albert Einstein to develop the particle model of light.

The photoelectric effect

In the photoelectric effect, a beam of incoming light shines on a metallic surface. When the beam hits the metal, it sometimes ejects electrons from the metal and sends the electrons down a tube to a collector. To do so, the beam must have sufficient energy to break the electrons' bonds to the metal and provide the electrons with sufficient kinetic energy to reach the collector. Reaching the collector requires a certain amount of minimum kinetic energy at emission, because an electric field exists between the collector and the emitter that acts to slow down the electrons on their path. See the figure below.



For now, focus your attention solely on the grayed tube at the top and ignore the portions of the circuit including the battery and ammeter. The photoelectric experiment allows us to test the wave model against the particle model, for this particular setup. As an experimenter, we have control

over both the intensity of the light and the color of the light. We can independently vary one variable or the other, and note the effect, enabling us to determine the appropriate model for this system.

So what might we see? The photoelectric effect can be explained using the conservation of energy. Light brings in a certain amount of energy. If the energy is sufficiently high, it frees an electron from the metal. Different metals bind the electrons with different amounts of energy, called the *work function*, and given the symbol W_0. If the incident light is less than the work function, the electrons remain attached to the plate.

Suppose the incident light has sufficient energy to free the electron from the plate. The electron emerges with a kinetic energy of at least 0 J, and possibly more. The energy of the incident light is split in some fashion between breaking the electrons bond to the metal and providing the electron with kinetic energy ($E_{\text{light}} = W_0 + KE_{\text{electron}}$). Higher energy light results in faster moving electrons.

Next, the electron travels from the emitter towards the collector. In this region, an electric field points from the emitter toward the collector. The electric force on the electron slows it down as it travels from the emitter to the collector. Thinking about energy again, the electron gains potential energy as it loses kinetic energy. As an experimenter, we we control the strength of the electric field, and thus the amount of potential energy the particle gains as it traverses the tube. We can measure the kinetic energy the electron had just after it was emitted from the first plate if we exactly stop the particle as it reaches the collector (this means we transfer exactly *all* of the electron's kinetic energy into potential energy). The potential required to do this is called the *stopping potential*. If we have a situation where many electrons reach the collector, we can slowly increase the voltage between the plates until we just reach the stopping potential.

When this experiment is run, we find the following:

- Different beam intensities have no effect on electron speed.
- Higher intensity beams free more electrons.
- Higher frequency beams result in electrons with higher speeds.
- Different frequencies have no effect on the number of freed electrons, provided the frequency is high enough that some electrons are freed.

These results all support the particle model of light. The intensity determines how many electrons are freed, because beams with higher intensities contain more photons. Intensity does not determine the speed of freed electrons because the energy of each photon in the light beam is determined by the frequency of the light, and changing the intensity does not change the energy of the beam.

The mathematics of the photoelectric effect

The previous section contains all of the concepts important to understanding the quantization of light. We now explore the mathematics to quantify these concepts. As you read, be sure to connect the equations to the concepts presented above. Our goal is to determine the energy of the incident light.

There are two main processes involved in the photoelectric effect. The first involves the light transferring energy to the electron, freeing it from the metal and giving it kinetic energy. Next, the electron travels down the tube, gaining potential energy and losing kinetic energy. As stated above, we adjust the voltage between the plates until the electron just barely stops short of the collector.

First, we will look at the second process, of slowing the electron down as it traverses the tube:

$$\Delta\text{PE} + \Delta\text{KE} = qV_{\text{final}} - qV_{\text{initial}} + \frac{1}{2}mv_{\text{final}}^2 - \frac{1}{2}mv_{\text{initial}}^2 = 0$$

We know that the final speed of the electrons is zero, since they just stop. Also, $V_{\text{final}} - V_{\text{initial}} = \Delta V$ is defined as V_{stopping}.We can rewrite our above equation as

$$\begin{aligned} qV_{\text{stopping}} + KE_{\text{initial}} &= 0 \\ \Rightarrow KE_{\text{initial}} &= -qV_{\text{stopping}} \end{aligned}$$

We know the charge of the electron ($q = -1.6 \times 10^{-19}$ C) and the stopping potential, so we have determined the kinetic energy of the electrons just after they emerge from the plate. Our goal is to relate this mathematically to the energy of the incident light. We can complete our task by recalling that the incident energy is transfered into breaking the electron off the metal with the remainder going into the electron's kinetic energy:

$$\begin{aligned} E_{\text{light}} &= W_0 + KE_{\text{initial}} \\ KE_{\text{initial}} &= E_{\text{light}} - W_0 \end{aligned}$$

We have called this kinetic energy "initial" to remind the reader that it is the kinetic energy that the electron has when it initially begins its trip toward the collector. Finally, we have found that:

$$-qV_{\text{stopping}} = E_{\text{light}} - W_0$$
$$E_{\text{light}} = W_0 - qV_{\text{stopping}}$$

If we find a way to adjust the energy of the light, then we will need to adjust the stopping potential, too, provided that we conduct our experiment with the same metal each time. If we do not need to adjust the potential, then we have not adjusted the energy of the light.

This setup is useful because the wave model and particle model hypothesize different ways of adjusting the energy of the light. We can try each method and note whether or not it resulted in a different electron kinetic energy (and thus different required stopping potential) to distinguish between the methods. Experimentally, we find the need that adjusting the frequency of the incident light requires us to adjust the stopping potential. This is the experimental evidence that the particle model accurately describes light in the photoelectric effect. Thus, $E_{\text{light}} = E_{\text{photon}} = hf$. Adding this relationship to our previous work,

$$hf = W_0 - qV_{\text{stopping}} > W_0 \qquad (\text{Recall } q < 0)$$

With this relationship, we could determine an experimental value for h, determine the work function of different medals, and more.

Which model is "correct?"

At this point, you might wonder which model is the "correct" model of light. The answer is "neither". In more sophisticated treatments we deal with light as a "quantum model" which incorporates all the examples we have discussed so far. We should refrain from saying that light is really this "quantum stuff" as future experiments may require us to replace this model with something else.

If neither model of light is correct, why do we teach them? Ultimately the full "quantum model" is just too difficult, and we can answer many questions by using the particle model or the wave model of light. Both of these simpler models correctly capture aspects of light's behaviour. Many books perpetrate confusion by claiming that light is somehow "both a particle and a wave", giving philosophers a lifetime of work. Many physicists are also guilty of

perpetrating this myth. We have a good "quantum model" for light (and electrons, and even whole atoms) – and in some situations we can simplify and use the wave model, while in others we can use the particle model. In other situations the "quantum model" does not fit into either a wave or particle description. Stuff on a microscopic scale behaves *differently* to our expectations, the fact that we cannot shoehorn it into our preconcieved notions does not mean quantum mechanics is contradictory. [5] It *does* mean that the microscopic world is highly counter-intuitive.

Because the wave-particle "duality" or "contradiction" is bought up so often, it bears repeating:

> Arguments about whether light is really a particle or a wave are a waste of oxygen, or worse yet, trees.[6] People making these statements are unaware of the concept of modelling and making approximations, and you should be hesitant to accept their advice about physics.

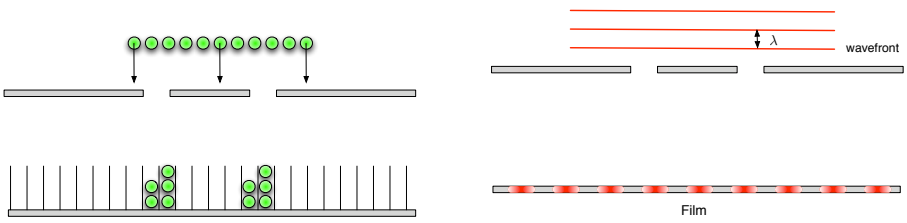
10-1-3 What are matter waves?*

10-1-3-1 Two-slit interference

Our definition for a wave has been too stringent – in unit 8 we referred to (material) waves as being oscillations of a medium about its equilibrium position in time and space. When we discussed light waves we no longer had an obvious medium, in fact light can travel through vacuum. The fact that light can reach us from the sun is a daily reminder of this fact. We still choose to refer to light as a wave because it obeys the principle of superposition. Superposition gives us constructive and destructive interference, and a dramatic example of this is two slit interference which we have seen in section 8-2-6. Light going through each slit superposes to give rise to the bright and dark bands as shown in figure 10-1.3 (b). Particles, on the other hand, can only go through one slit at a time and cannot interfere. The pattern particles would make is shown in figure 10-1.3 (a). To compare the two we consider the "number of marbles in a bin" to be analogous to the brightness of the light.

[5] There are other questions related to the probabilistic interpretation of quantum mechanics, and whether or not that ultimately "makes sense". Most physicists are convinced it does.

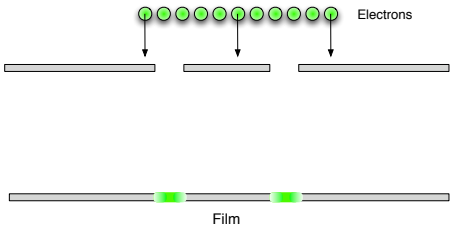
[6] We are completely aware of the irony.



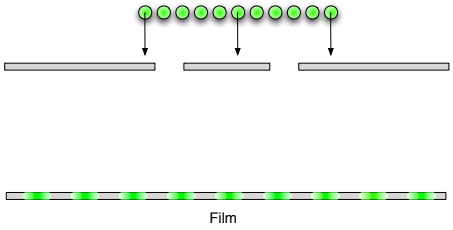
(a) Marbles through two-slits (b) Waves through two-slits

Figure 10-1.3: We see that the particles (such as marbles) produce a different pattern than waves (such as waves) would. For the marbles, we replace the idea of "brightness" with the number of marbles in a given bin.

What would electrons do if they were also passed through two slits? Because we are used to thinking of electrons as little balls, we may expect to get a "particle-like" interference pattern:



When this experiment is actually performed, we get quite an unexpected result – the electrons form an interference pattern similar to waves:



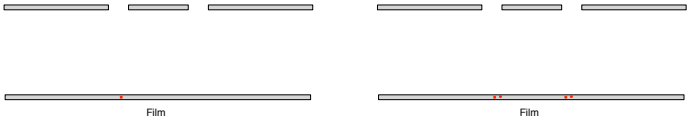
In fact, the interference pattern created by the electrons is identical to one created by light with the same phase difference between the two slits and a wavelength

$$\lambda = \frac{h}{p} \approx \frac{h}{mv} \qquad (\text{if } v \ll c).$$

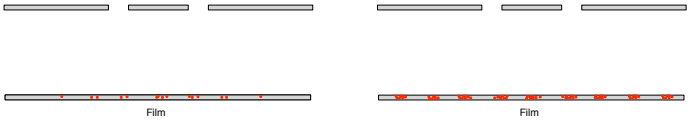
This is called the *de Broglie wavelength*. It is the same formula we learned for photons, except there we wanted the momentum: $p = h/\lambda$. Just as light can exhibit particle-like properties (such as in the photoelectric effect), matter can exhibit wave like effects (such as interference patterns).

10-1-3-2 Single electron and single photon interference

We learned when studying waves the reason that the interference pattern occurs is because we have a superposition of the wave that goes through each slit. What would be the pattern we would expect if we sent through one electron at a time, and kept track of where they land? Or, for that matter, one photon at a time? We will use a film to keep track of where the photons land, and let many photons through our slits one at a time. As both the photon and the electron are supposed to be a single indivisible unit, and should go through one slit or another. We may expect to recover the "marble-like" interference pattern again. For concreteness, let us discuss the experiment for photons – exactly the same analysis works for electrons as well. We show the film after sending through a single photon (left diagram) and then four photons.

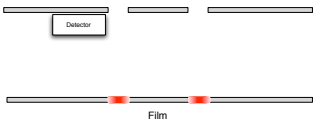


So far so good. As we continue sending photons through, we see that the pattern we build up was *not* just the "marble" pattern shown in figure 10-1.3 (a). Instead the pattern that builds up resembles that of a wave:



Somehow the individual photons going through the slit have managed to go through both slits and produce an interference pattern. The light is still "particle-like" in the sense that each photon makes only one bright spot on the film (as in the photoelectric effect) but produces a pattern that we expect waves to make.

It seems like the particle (be it photon or electron) is travelling through *both* slits and interfering with itself! Because photons and electrons are supposed to be indivisible, this seems like a ridiculous statement. To get further insight on this, let us set up a detector by one of the slits that tells us if the photon goes through that slit. That way we should be able to eliminate the possibility that the photon is "interfering with itself". If we try this experiment with the detector, the result we obtain on our film is



Now we can tell which photon went through which slit, but we get the interference pattern we expect from particles. By switching the detector off again, we recover the interference pattern for waves. Another way of putting the same thing is that when we tried to probe the particle nature of the light (e.g. which slit did it go through) we ended up with the pattern we expect marbles to make. When we ignored which slit it went through, somehow the photon went through both and interfered with itself. In either case, *how the light acts depends on what we are measuring!* We can no longer simply ignore the influence of our measuring apparatus on the system, a subject we will address again in section 10-1-3-6.

There is another fundamental issue which has not been addressed: what distinguished different photons and made one make one spot on the film, and another photon make a different spot? *Nothing* distinguishes these different photons. Unlike the marbles that were shown in figure 10-1.3 (a) where marbles starting in different places ended in different places these photons are identical.

Put in a slightly different way, given a photon we *cannot* predict where it will land. It is much more likely to end up in a constructive interference fringe (after all, the reason they are "bright" is that many photons go there) and has no probability of ending up where there is complete destructive interference. But we have given up deterministic physics – we must now settle with being able to calculate *probabilities* instead.

10-1-3-3 Interpreting the wave

We have stated that this "matter wave" will give rise to an interference pattern. But what else does the wave tell us? The interference pattern gives us a clue: the "brighter" the spot on the screen (for the full interference pattern) the more likely it had to be for the particle to land there. We generalise this to

$$P\left(\text{particle in region around } x \pm \frac{\Delta x}{2}\right) \propto \text{energy dumped in that region of screen by resulting fringes}$$

The energy dumped into the screen for light is $I\,\Delta x$. This is true provided Δx is small, so that the intensity is roughly constant. If we want to know about a large region we should use an integral instead. The intensity I is

proportional to the *square* of the field. We then propose that we will get the right answer if:

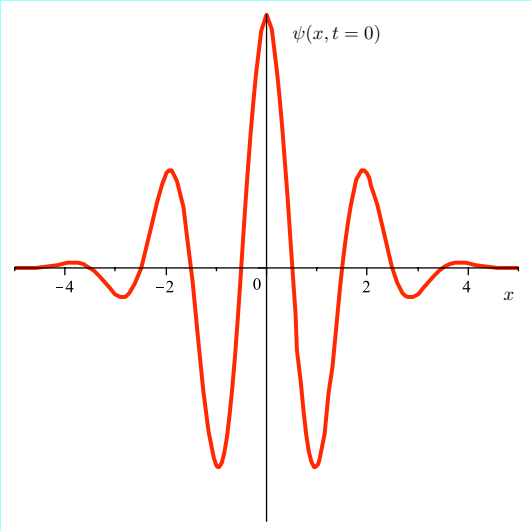
$$P\left(\text{particle in region around } x \pm \frac{\Delta x}{2}\right) \propto |\psi(x,t)|^2\,\Delta x$$

where $\psi(x,t)$ is the wave function for this matter wave. This is not a derivation, as quantum mechanics is a completely new phenomena that cannot be derived from our discussion of electromagnetism or Newton's laws. Instead it is an argument to make the choice of $|\psi|^2$ seem less *ad hoc*. This choice must be (and has been) verified by experiment to be taken seriously.

A summary of the main idea above is that the results of experiments are now only determined up to probabilities. If we plot $|\psi(x,t)|^2$ against x, the probability of a particle being in a particular region is proportional to the area of that region. An example should clarify the main points.

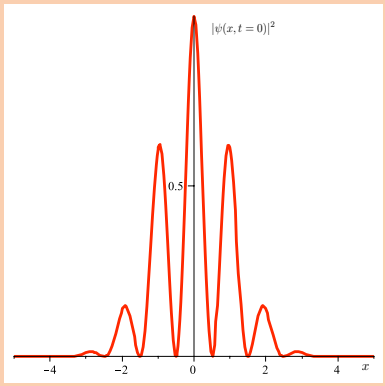
Example #6:

Below is the "matter wave" for the electron. Which location or locations is the electron most likely to be found? Which location or locations is the electron least likely to be found?

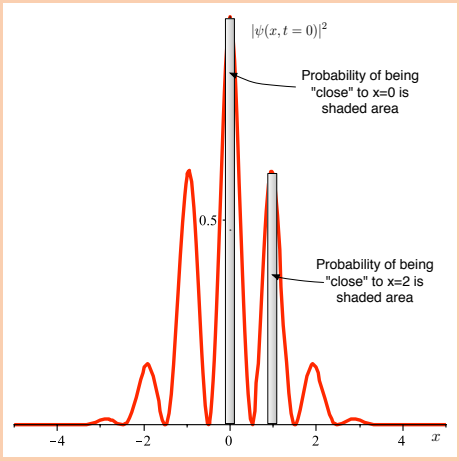


Solution:

We first want to turn this graph into a probability distribution, so we take the square of this graph. This gives us the graph below:



Now we need to figure out what the question is actually asking. This graph peaks at $x = 0$. Now the probability that the particle is at *exactly* $x = 0$ is *zero* – the area under this point is tall but has basically zero width. However, the probability that the particle is *around* $x = 0$ is quite a bit higher that around any other point. For example, compare the area around $x = 0$ to the area around $x = 2$ – we see the particle is almost twice as likely to be close to $x = 0$, and certainly more likely than being around $x = 3$.



However, it is true that it is more likely for the particle to be between $x = 0.5$ and $x = 2.5$ than "near" $x = 0$, as there is more area under the curve. Taking this to the extreme, the probability must be one that the particle is somewhere, so the region with the most probability is the one that extends from $-\infty$ to ∞!

The question was asking which *location* was the particle most likely to be found, not which *region* was the most likely for the particle to be found. If we look in a tiny region around $x = 0$ it has the most area of any other

"tiny region", so we call $x = 0$ the most likely location for the particle. So while strictly the probability of being at any particular location is zero, the statement "the particle is most likely at $x = 0$" is really a convenient shorthand for "the particle is more likely to be very close to $x = 0$ than be very close to any other single point". It is admittedly a sloppy use of language, but provided you know what is meant by such statements it will not lead you astray. The actual numerical value for the probability depends on how large we make this "tiny region".

The least likely locations for the particle is whenever the probability wave goes through zero. These locations are $x = \pm 0.5$, $x = \pm 1.5$, $x = \pm 2.5$ and pretty much anything with $|x| > 3.5$.

10-1-3-4 Matter waves: the trouble with frequency and wave speed

Matter waves are analogous to the other waves we have seen so far in this course; the major difference is that matter waves have *no polarisation*. The oscillations in a probability wave do not have a sensible interpretation as being in a direction so the concepts of longitudinal or transverse polarisations simply don't apply.

More problematic is the issue of frequency and wave speed. When we discussed waves in unit 8 we considered the frequency f and related it to the wave speed by $v_{\text{wave}} = \lambda f$. To carry on the analogy with light, we may think that for a particle $E_{\text{particle}} = hf$. So far all of this is true, but we have glossed over one important detail: we don't know v_{wave} for the matter wave! In particular, it is *not* the same as the speed of the particle!

$$v_{\text{wave}} \neq v_{\text{particle}}$$

For light $v_{\text{wave}} = v_{\text{photon}} = c$. For matter waves, the velocity is much more complicated and at this level you should avoid thinking about the frequency of matter waves[7]. Instead, we shall mean v_{particle} when we write v, and simply restrict our attention to the wavelength $\lambda = h/p$.

[7]Technical aside: you may wonder how the velocity of a particle and the velocity of its wave can ever be different. Waves have many velocities, and particles travel at the *group velocity* of matter waves. The velocity v_{wave} is the *phase velocity* and is unobservable. For light in vacuum these velocities are both c, so no confusion can result. (Phase velocity and group velocity are given so you can read up on this if you are interested, they are not important for this course.)

When we looked at a particle in a harmonic oscillator, we will saw f in the formula for the energy levels. This f does not correspond to the frequency of the matter wave f_{wave}, but refers to the frequency of the particle f_{particle} as it bounces back and forth. Because both the particle and the wave are oscillating this can cause some confusion. In this class the details are far too technical, and we will *never* look at the frequency of the matter wave. Whenever we write v or f in this section, we are always referring to v_{particle} or f_{particle} respectively.

10-1-3-5 Using matter waves to find the energies: an example

We have simply given you the formula for the allowed energy levels E_n in the infinite well, the simple harmonic oscillator and the hydrogen atom. Now we will show you how to actually *find* the energy levels in the case of the infinite square well because it is the simplest to do, and indicate why the other cases are slightly harder. If you need a reminder on the infinite square well, flip back to page 209.

Because the potential energy inside the well is zero, the total energy of the particle E is simply equal to the kinetic energy of the particle:

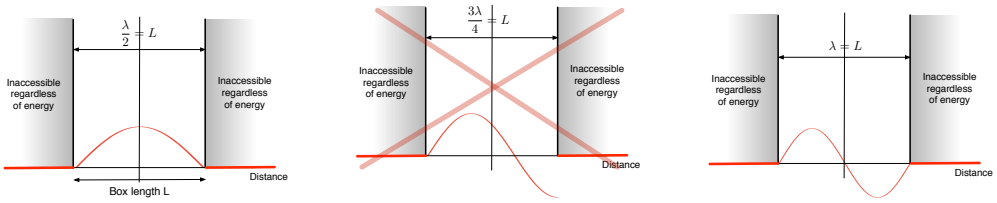
$$E = \text{KE} + 0 = \frac{1}{2}mv^2$$

Because the total energy is not changing, the speed cannot change, and the magnitude of the momentum ($|p| = m|v|$) cannot change either. Because the momentum and wavelength are related by

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

we see that the matter wave has the same wavelength throughout the well.

We also know that the particle cannot escape the box, so there is *zero* probability the particle is located outside the box. This tells us that the matter wave had better be zero outside the box. Because the matter wave should not undergo sudden jumps, this restricts which wavelengths we are allowed. Here are three examples of matter waves that have zero amplitude outside the box, and constant wavelength inside:



Notice that the first and last example work well, but in the second example the matter wave "jumps". The wavelength corresponding to this second example is simply *not allowed* in this system. The only wavelengths that are allowed are those that go to zero on the two walls. But this is *exactly the same* as standing waves with both of their ends fixed, which we discussed in §8-2-5. The allowed wavelengths are then

$$\lambda_n = \frac{2L}{n}, \qquad n = 1, 2, 3, \ldots$$

where n is the number of *anti-nodes* in the wave.

Because we know the momentum is related to the wavelength, restricting the wavelengths to certain values also restricts the allowed values of momentum:

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

We know that $p = mv$, and because the mass does not change we see that quantising momentum aslo has the effect of quantising the velocities:

$$v_n = \frac{p_n}{m} = n\frac{h}{2Lm}.$$

Finally, because we are only allowed certain values of the velocity we can only have certain values for the kinetic energy:

$$E_n = \frac{1}{2}mv_n^2 = n^2\frac{h^2}{8mL^2} = n^1 E_1$$

which is exactly the formula we found before. But now we have some idea of what n is and how it got there: n is counting the number of anti-nodes in our matter wave!

Let us summarise the process:

$$\begin{aligned} \text{Standing probability waves} &\Longrightarrow \text{quantized } \lambda \\ \text{Quantized } \lambda &\Longrightarrow \text{quantized } p \\ \text{Quantized } p &\Longrightarrow \text{quantized } v \\ \text{Quantized } v &\Longrightarrow \text{quantized KE} \end{aligned}$$

Other potentials are more difficult to calculate exactly. One reason for this is the potential energy keeps changing, so the kinetic energy (and therefore wavelength) keep changing. However, with some more mathematical and physical consideration, it is possible to calculate the spectrum E_n of other potentials.

10-1-3-6 Uncertainty principle

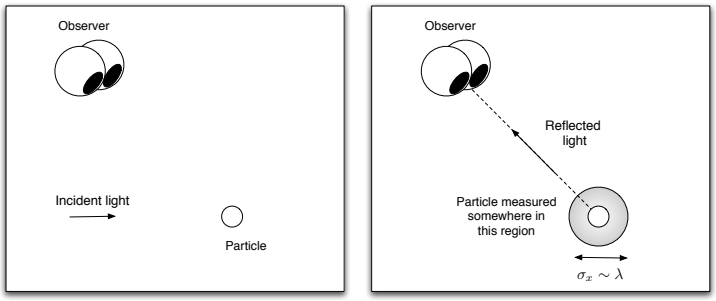
The idea that a measurement *inevitably* affects the particle being observed has already been introduced when looking at the two slit experiment. In that situation we discovered that trying to measure which path the particle took resulted in the interference pattern disappearing, and the "marble-like" interference pattern reappearing.

Measurement affects physical systems in more general ways. One very famous example is the *Heisenberg uncertainty principle* which states that you cannot know both the position and momentum of particle arbitrarily well. If you know the particle is in a region of size σ_x, and you know the momentum within σ_p then the following inequality must be satisfied[8]

$$\sigma_x \, \sigma_p \geq \frac{h}{4\pi}.$$

The consequence of this is as we try and get a more accurate measurement of position ($\sigma_x \to 0$) the information we have about the momentum gets worse ($\sigma_p \geq h/(4\pi\sigma_x)$).

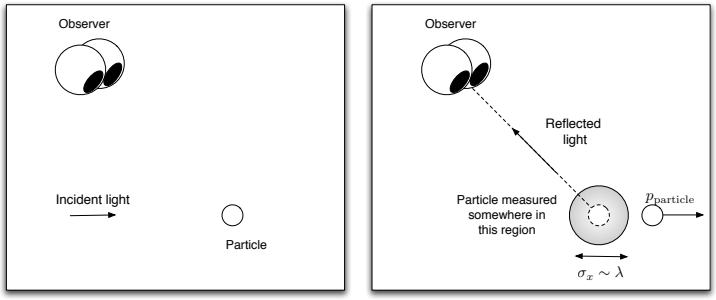
One way of seeing why something like this should be true is to consider the actual act of measuring the position of a particle. One way of doing this would be to shine light on it, and see where the goes (i.e. trace back the rays as we did for lenses). This is similar to how we see objects in everyday life:



[8]Technically σ_x and σ_p are the standard deviations in measurements of the particle's position and momentum respectively, rather than the region we "absolutely know" they must be in.

We cannot determine the position of the particle to a size smaller than the wavelength of light we use, so the error involved in the measurement of the particles position is $\sigma_x \approx \lambda$. As we use light of shorter and shorter wavelengths we get better accuracy on the particle's position.

As the light has momentum, it will "kick" the particle as it changes direction after interacting with the particle:



Because momentum is conserved, the amount of momentum transferred to the particle depends on how much the light was deflected by. If the light is only slightly deflected, the particle will still have (approximately) zero momentum. If the light bounces directly back, the momentum of the particle must be *twice* the initial momentum of the photon $2h/\lambda$. The uncertainty in the final momentum of the particle is roughly $\sigma_p \sim \frac{1}{2}(p_{\max} - p_{\min}) = h/\lambda$. Putting these together we have

$$\sigma_x \, \sigma_p \approx (\lambda) \left(\frac{h}{\lambda} \right) = h \tag{10-1.2}$$

We have been rough in our estimates, but this shows us that by using light of a long wavelength we can measure momentum well ($p_{\text{light}} \approx 0$) but know almost nothing about where the particle is **or** sacrifice accuracy in our knowledge of momentum for more accuracy in position.

An obvious objection to the above argument is that there may be other ways of measuring position and momentum that do not involve light, and that all that has been demonstrated is this particular method cannot accurately measure both simultaneously. This is a very valid objection, but it turns out that *any* measuring procedure will disturb the system in such a way that you cannot simultaneously determine the position and momentum completely.

Hidden varaibles?
A more fundamental flaw with this derivation is that it pretends the particle is really a "marble-like" particle, and the uncertainty in position or momentum comes about solely from our lack of ability to measure it. This is the idea that there are things hidden away that we have not been clever enough to measure, collectively called "hidden variables".

Quantum mechanics actually makes a much stronger statement: the uncertainty in measurements does not reflect our inability to make certain measurements but rather the particle does not really have a particular position or momentum until one makes a measurement. This may seem like a philosophical distinction, how could one show that these quantum particles were not really "marble-like" and that all the uncertainty came from our inability to measure? On the other side if we cannot measure it, can we really talk about it being true? One of the greatest triumphs is that quantum mechanics is strange enough that there *are* experimental differences between "hidden variables" where the uncertainties are just from bad measurements, and quantum mechanics. In all the experimental tests the predictions of quantum mechanics have been borne out.[9]

What is the take home message from this? The argument using light to measure the position of a particle to demonstrate why the uncertainty principle is true is one that appeals to our sense of how the world should work at the level of tennis balls and things we are used to. However, it suffers major logical flaws e.g. how are we sure this is the best measurement that can be performed? Quantum mechanics makes a much more radical departure: it claims that the world is not like tennis balls at all! In retrospect, it is somewhat amusing that the tennis ball like argument above gets us so close to the right answer!

10-1-4 Summary

We have seen that at the microscopic level the world is very different from anything we are familiar with. Yet the world we are familiar with is built out of this strange quantum world. Quantum mechanics brings three radically different ideas to physics:

[9]These experiments are too subtle to discuss in this course, but if you are interested in them look under *Bell's inequalities*.

1. *Some continuous quantities are quantised*
   It is perhaps easiest to understand why this one is not noticeable in everyday life, as the steps between allowed values are typically quite small (c.f. water molecules vs. continuous water).

   - Light could be described by "packets" of energy called photons. The energy of a photon is $E = hf$, where h is a universal constant known as Planck's constant. A photon has momentum given by $p = E/c$.
   - The information needed to label a state are known as the *quantum numbers*. In the examples above, n is a quantum number.
   - In addition to the shell number n, electrons in atoms have three additional quantum numbers, all of which are to do with angular momentum.
     - ℓ: takes values between 0 and $n-1$. Labels how much angular momentum the electron has going around the nucleus.
     - m_z which lies between $-\ell$ and ℓ carries information about how much angular momentum is in the z-direction.
     - m_s which takes two values. This is to do with the electron's internal spin.
   - The Pauli exclusion principle forbids any two electrons from sharing the same quantum state. Once accepted this principle tells us why the electrons fill up the states of an atom, rather than crowding into the low energy states. The distribution and range of quantum numbers is largely responsible for the structure of the periodic table.

2. *The results of measurements are no longer deterministic, but instead have different probabilities*
   For systems with large number of particles this becomes unimportant, as the probability of being a long way from the mean value is small. The source of the probability is not due to imperfect knowledge as it would be for flipping a coin or roulette[10], but is instead due to fundamental randomness.

   - Objects we normally consider to be particles have a wavelength given by $\lambda = h/p = h/(mv)$.

[10]See the book *Bringing down the house* about six M.I.T. students that measured the velocity and position of a ball to win against the casino. In a quantum game, this sort of cheating would be impossible.

- The absolute square of the wavelength gives the probability density for the particle being observed at a particular location.
- Confining a particle to a small region leads to quantisation of energy. We discussed the *explicit* allowed energy levels for three different systems. For each of these systems n is a positive integer.

3. *The act of measuring a system affects it in a substantial way.*

In special circumstances we can get quantum mechanical affects at large scales – the primary example is *superconductivity*.

Finally, we summarise the actual systems we explicitly gave the energy levels for in the table below:

System	**Formula for energy levels**
Particle of mass m in an (unescapable) box L	$E_n = n^2 E_1 = n^2 h^2/(8mL^2)$
Simple harmonic oscillator, classical frequency f	$E_n = \left(n - \frac{1}{2}\right) hf$
A single electron atom in an element with Z protons	$E_n = -Z^2(2\pi)^2 mk^2 e^4/(2h^2)$ $= -Z^2(13.6\,\text{eV})/n^2$

Appendix A: Physical constants

Memorizing these constants is not doing physics. They are provided for your convenience, but we will give you values for constants on quizzes and finals. You should memorize the SI unit conventions, and should know the "approximate values".

Fundamental constants

Quantity	Symbol	Value
Speed of light	c	3.00×10^{8} m/s
Gravitational constant	G	6.67×10^{-11} Nm^2/kg^2
Planck's constant	h	6.63×10^{-34} J s
		4.14×10^{-15} eV s
Coloumb's constant	k	9×10^{9} N m^2/C^2
Permeability of the vacuum	μ_0	$4\pi \times 10^{-7}$ N s^2/C^2
Avagadro's number	N_A	6.023×10^{23} atoms/mole
Boltzmann's constant	k_B	1.38×10^{-23} J/K
Stephan-Boltzmann constant	σ	5.67×10^{-8} W m^{-2} K^{-4}
Proton charge	e	1.602×10^{-19} C

Matter

Particle	Mass	Charge
Electron	9.11×10^{-31} kg	-1.60×10^{-19} C
Proton	1.67×10^{-27} kg	1.60×10^{-19} C
Neutron	1.67×10^{-27} kg	0 C

Optics

Different colours have slightly different refractive indices. This table has approximate values for the visible spectrum (exact values will depend on the exact material and frequency)

Material	$n = c/v_{\text{medium}}$
Vacuum	1.0 (exact)
Air	1.0003
Water	1.33
Glass (crown)	1.50–1.62
Glass (flint)	1.57 – 1.75
Silicon	3.5
Germanium	4.0
Diamond	2.42
Eye	1.33
Eye lens	1.41

Approximate values

These figures give you a *rough* idea of how big various quantites are, such as the well-depth of the Lennard-Jones potential. Exact values depend on the system being considered.

Quantity	Approx. Value	
Size of atom	$\sim 10^{-10}$ m	1 Å
Well-depth	10^{-21} J	10^{-3} eV
Ionization energy	10^{-20} – 10^{-18} J	0.1 – 10 eV
Mass of an atom	10^{-27} – 10^{-25} kg	1 – 200 amu
Visible light :		
frequency	6×10^{14} Hz	—
E	3×10^{-19} J	2 eV
Tallest building	508 m	
Height of Everest	8850 m	
Radius of Earth	6380 m	

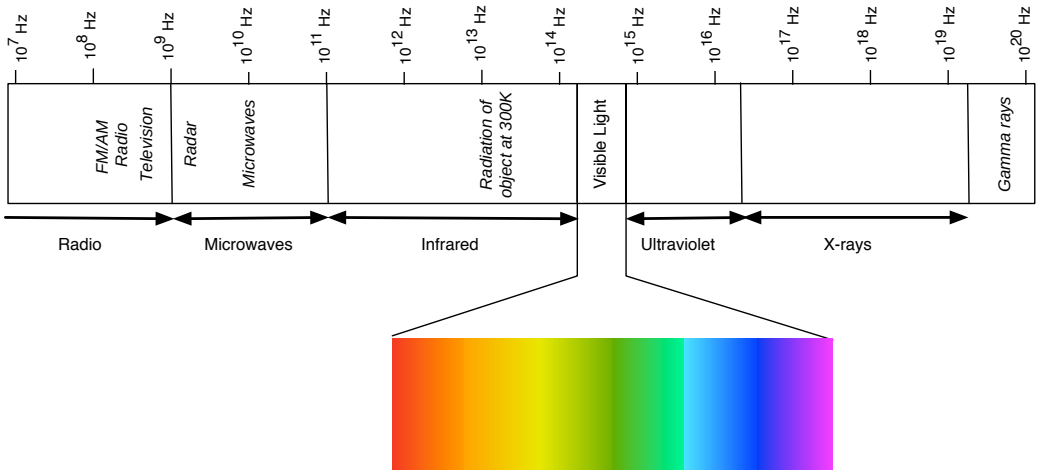
Solar system data

Body	Mass	Radius	Distance from sun	Orbital period	Surface temp (ave)
Sun	2×10^{30} kg	695000 km	—	—	6000 + 273 K
Mercury	3.3×10^{23} kg	2400 km	5.7×10^{6} km	88 days	179+273 K
Venus	4.9×10^{24} kg	6050 km	1.1×10^{9} km	225 days	482+273 K
Earth	5.98×10^{24} kg	6380 km	1.5×10^{9} km	365.25 days	15 + 273 K
Moon	7.3×10^{22} kg	1737 km	–	27.3 days	-46 + 273 K
Mars	6.4×10^{23} kg	3400 km	2.3×10^{9} km	687 days	-63 + 273 K
Jupiter	1.9×10^{27} kg	71500 km	7.8×10^{9} km	4332 days	-121 + 273 K
Saturn	5.7×10^{26} kg	60300 km	1.4×10^{10} km	29.5 years	-125 + 273 K
Uranus	8.7×10^{25} kg	25600 km	2.9×10^{10} km	84 years	-193 + 273
Neptune	1.0×10^{26} kg	24746 km	4.5×10^{10} km	164.8 years	-185 + 273 K

One day means 23.9345 hours. Years are all Earth years. Earth-Moon distance is 384000 km.
$G = 6.67 \times 10^{-11}$ N m^2 kg^{-2} is useful in calculating $\boldsymbol{g}$ at the surface of these bodies.

Electromagnetic spectrum

Light is often characterized by wavelength. This is incorrect; red light always has the same frequency but the wavelength depends on the index of refraction (and hence the medium). When someone says that "red light has a wavelength of 700 nm" this is understood to be the wavelength in vacuum.



Colour	λ_{vac} range	λ middle	Frequency	Energy
Red	620–750 nm	700 nm	4.3×10^{14} Hz	1.8 eV
Orange	590–620 nm	600 nm	5.0×10^{14} Hz	2.0 eV
Yellow	570–590 nm	580 nm	5.1×10^{14} Hz	2.1 eV
Green	495–570 nm	540 nm	5.5×10^{14} Hz	2.3 eV
Blue	450–495 nm	470 nm	6.4×10^{14} Hz	2.6 eV
Violet	380–450 nm	400 nm	7.5×10^{14} Hz	3.1 eV

Part of Spectrum	**Typical size in vacuum**
Short wave radio	$\lambda \sim$ Building
AM/FM/TV	$\lambda \sim$ Person
Microwaves	$\lambda \sim$ Insect
Infrared	$\lambda \sim$ Flea
Visible	$\lambda \sim$ Cells
Ultraviolet	$\lambda \sim$ Molecules
X-rays	$\lambda \sim$ Atoms
γ-rays	$\lambda \sim$ Nuclei

Units

You should memorize what the SI prefixes mean. They are used in all branches of science, and can be given on quizzes and the final without explaining what they mean.

SI prefixes

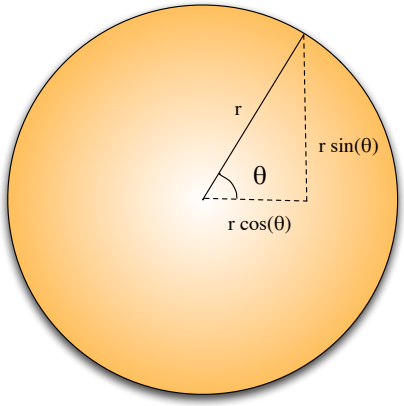
Name	**Symbol**	**Meaning**	**Example**
femto	f	$\times 10^{-15}$	1 fm = 10^{-15} m
pico	p	$\times 10^{-12}$	1 pm = 10^{-12} m
nano	n	$\times 10^{-9}$	1 nm = 10^{-9} m
micro	μ	$\times 10^{-6}$	1 μm = 10^{-6} m
milli	m	$\times 10^{-3}$	1 mm = 10^{-3} m
kilo	k	$\times 10^{3}$	1 km = 10^{3} m
mega	M	$\times 10^{6}$	1 Mm = 10^{6} m
giga	G	$\times 10^{9}$	1 Gm = 10^{9} m
tera	T	$\times 10^{12}$	1 Tm = 10^{12} m
peta	P	$\times 10^{15}$	1 Pm = 10^{15} m

Common non-SI Units

Non-SI Unit	**Measures**	**Conversion to SI**
Angstom Å	Length	1 Å= 10^{-10} m
Electron-Volt	Energy	1 eV = 1.602×10^{-19} J
Atomic mass unit	Mass	1 amu = 1.66×10^{-27} kg

Appendix B: Trigonometry

Circle



$$\sin A + \sin B = 2 \sin\left(\frac{A+B}{2}\right) \cos\left(\frac{A-B}{2}\right)$$
$$\cos A + \cos B = 2 \cos\left(\frac{A+B}{2}\right) \cos\left(\frac{A-B}{2}\right)$$

$$\sin(A+B) = \sin A \cos B + \sin B \cos A$$
$$\cos(A+B) = \cos A \cos B - \sin A \sin B$$
$$\tan(A+B) = \frac{\tan A + \tan B}{1 - \tan A \, \tan B}$$

$$\sin\theta \approx \theta, \quad \theta \text{ in radians}$$
$$\tan\theta \approx \theta, \quad \theta \text{ in radians}$$

Appendix C: Vectors

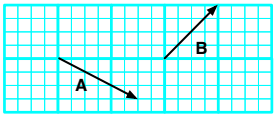
A *vector* is a quantity with a magnitude and direction, and in these notes will be denoted by a bold letter (such as $\boldsymbol{A}$). We represent a vector by a straight arrow; the length of the arrow represents the vectors magnitude, while the direction is given by the direction the arrow points in.

Adding vectors

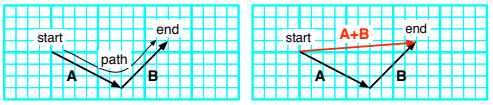
Vectors do not add like numbers!
It is often useful to combine vectors by vector addition. For example, if we are looking for the total momentum of a system we add all the momentum vectors. To find the net force on an object, we add all the forces acting on that object together. To find the total electric field at a particular location, we add all the electric fields at that location together. While these examples all refer to different physical situations, the way vectors are added together is the same.

Graphical addition

The grid below shows two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$.



To add these vectors we join the arrows up to make a "path" which we can follow, *always going in the direction of the arrows*. The vector $\boldsymbol{A} + \boldsymbol{B}$ is then the vector that connects the beginning of this path to the end.



Components

Another method for adding vectors is by breaking a vector up into components. While vectors don't add like numbers, the components of a vector do. We will break this vector into x- and y-components, which are the most common choice. To do this we must figure out how many "units" the vector points right and how many "units" the vectors points up. Sometimes we have a grid like in the example above in which case we can just count the number of units, but in many situations we shall have to use trigonometry to break a vector into components. In the example above:

$\boldsymbol{A}$: **6** units right, **−3** units up
$\boldsymbol{B}$: **4** units right, **4** units up
$\boldsymbol{A} + \boldsymbol{B}$: **10** units right, **1** unit up

A common mistake

The magnitudes of vectors do not add like numbers. Notice that in the example given, the length of the three vectors can be found by Pythagorus's theorem.

The lengths are:

$$\begin{aligned} |\boldsymbol{A}| &= \sqrt{(\mathbf{6})^2 + (\mathbf{-3})^2}\,\text{units} \\ &= 6.71\,\text{units} \\ |\boldsymbol{B}| &= \sqrt{(\mathbf{4})^2 + (\mathbf{4})^2}\,\text{units} \\ &= 10.05\,\text{units} \\ |\boldsymbol{A}+\boldsymbol{B}| &= \sqrt{(6+4)^2 + (-3+4)^2}\ \text{units} \\ &= 10.05\,\text{units} \\ &\neq 6.71\,\text{units} + 5.66\,\text{units} \end{aligned}$$

The length of a vector $\boldsymbol{A}$ is denoted $|\boldsymbol{A}|$.

Subtracting vectors

The negative of a vector $-\boldsymbol{B}$ is $\boldsymbol{B}$ with the arrow pointing in the opposite direction. To subtract $\boldsymbol{B}$ from $\boldsymbol{A}$, we *add* vectors $\boldsymbol{A}$ with the vector $\boldsymbol{-B}$:

$$\boldsymbol{A} - \boldsymbol{B} = \boldsymbol{A} + (-\boldsymbol{B})$$

Multiplying vectors

(Called "scalar multiplication" in other texts)
Multiplying a vector by a positive number changes the magnitude of a vector, but not the direction. For example $2\boldsymbol{A}$ points in the same direction as $\boldsymbol{A}$ but is twice as long.

Multiplying a vector by a negative number changes the magnitude of the vector and makes it point in the *opposite* direction. For example $-2\boldsymbol{A}$ points in the opposite direction as $\boldsymbol{A}$ and is twice as long.

If we multiply a vector by a number with units, the final vector has a magnitude with these new units as well. An example that shows all of these possibilities is

$$\boldsymbol{F}_{\text{field on } q} = q\boldsymbol{E}$$

where

- The units of force are N, of q are C and of $\boldsymbol{E}$ are N/C.
- If q is positive, $\boldsymbol{F}_{\text{field on } q}$ and $\boldsymbol{E}$ are in the *same* direction.
- If q is negative, $\boldsymbol{F}_{\text{field on } q}$ and $\boldsymbol{E}$ are in *opposite* directions.

Into and out of the page

Drawing vectors going into the page or out of the page is difficult, so a special notation has been adopted for this purpose. The symbols we use to represent a vector going in or out of the page are chosen because they look like a dart going in or out of the page:

Representing a vector coming out of the page

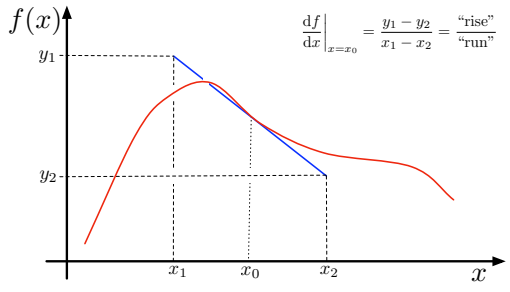
Representing a vector going into the page

Warning

This was a brief recap of vectors that should be familiar to you from 7B. If you still find vectors confusing, look through your 7B course notes or ask a TA during office hours.

Appendix D: Calculus

Differentiation



This course is aimed at teaching concepts, but some advanced mathematics is required. We want you to be able to "graphically differentiate" functions. This means identifying the tangent line at a particular point, and finding the slope of the tangent line using

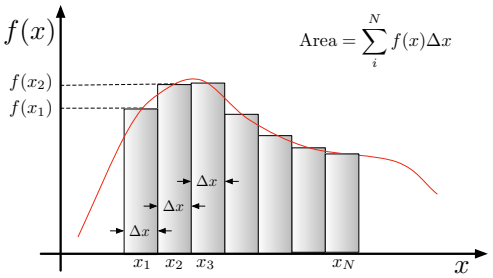
$$\text{slope} = \frac{\text{“rise”}}{\text{“run”}} = \frac{\Delta y}{\Delta x}.$$

The graph above shows an example of finding the tangent line (in blue) and calculating the slope of the function at the point x_0. The table below gives some uesful derivatives (only required for some professors):

Some useful derivatives	
$f(x)$	$f'(x)$
Ax	A
$A\sin(\omega x + \phi)$	$A\omega\cos(\omega x + \phi)$
$A\cos(\omega x + \phi)$	$-A\omega\sin(\omega x + \phi)$
$1/x$	$-1/x^2$

A, ω and ϕ are all constants in the above.

Integration



In this class we will be interested in quantities accumulated from the initial to the final point, represented by area under the curve. An approximation to the area under the curve between x_1 and x_N is given by the area in the shaded rectangles, each with width Δx. As these rectangles become *infinitely thin* the answer becomes exact. This exact value is called the *integral* of the function.

In this course you will be expected to know some basic facts about integrals:

$$\int_{x_i}^{x_f} Af(x)\,\mathrm{d}x = A\int_{x_i}^{x^f} f(x)\,\mathrm{d}x \qquad A \text{ any constant}$$

$$\int_{x_i}^{x_f} \mathrm{d}x = \Delta x$$

$$\int_{x_i}^{x_f} x\,\mathrm{d}x = \frac{1}{2}(x_f^2 - x_i^2)$$

Some useful integrals for doing physics,

but not *essential* for this class, are

$$\int_{x_i}^{x_f} \frac{1}{x^2}\,\mathrm{d}x = -\frac{1}{x_f} + \frac{1}{x_i}$$

$$\int_{x_i}^{x_f} \frac{1}{x}\,\mathrm{d}x = \ln\left|\frac{x_f}{x_i}\right|$$

$$\int_{x_i}^{x_f} \cos(\omega x + \phi)\,\mathrm{d}x = \frac{1}{\omega}\Big(\sin(\omega x_f + \phi) - \sin(\omega x_i + \phi)\Big)$$

$$\int_{x_i}^{x_f} \sin(\omega x + \phi)\,\mathrm{d}x = -\frac{1}{\omega}\Big(\cos(\omega x_f + \phi) - \cos(\omega x_i + \phi)\Big)$$

If you are good with mathematics these integrals are good for seeing interesting relationships. The focus of the course is conceptual and not mathematical, so do not spend a lot of time trying to memorise these integrals.

Summary

While this course is not mathematically based, one of the course prerequisites is Math 16B. As it is a prerequisite no time on class will be spent covering this material. If you are uncertain of any of this material you should review your calculus text.

Index