PHYSICS 7C
GENERAL PHYSICS

Dina Zhabinskaya
UC Davis

UC Davis

Physics 7C

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

8: Waves

8.1: 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 the *harmonic plane wave model*, which we will use extensively to model and understand a wide variety of wave phenomena.

In this section we will familiarize ourselves with waves by focusing on *material waves*. These are the disturbances of atoms or molecules in a particular substance. Any sort of ripple that you have seen is a material wave. Some examples include ripples in a pond or a wave on a slinky or a string. Although not visible to the naked eye, sound is modeled by material waves as well.

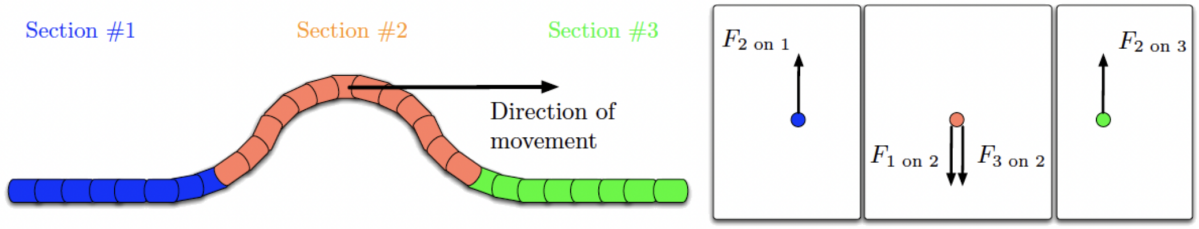
One of the distinguishing features of physics is that physicists continually strive for general principles and simple models that can be applied to large classes of phenomena. In our study of wave phenomena we consciously take this approach. The focus is on the models and their representations, not on any one of an almost unlimited number of individual examples associated with waves (like 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 apply to any phenomenon that can be modeled as a wave.

Basic Wave Concepts

There are two important goals associated with the first part of this unit. Firstly, to become familiar with wave phenomena and how we analyze them, and secondly, to sufficiently understand the mathematical representation of *one-dimensional harmonic waves*. We want to use this mathematical representation as a tool throughout the rest of the course to help us understand the physics of sound, light, and other types of waves.

A material wave is a very common type of internal motion of a material substance. That substance through which the wave propagates is called the *medium*. In order for material waves to exist there must be forces between neighboring particles in the medium. For simplicity will examine how a disturbance travels and gives rise to wave-like behavior by coloring a medium three different pieces and labeling them as sections 1, 2 and 3 in Figure 8.1.1 below.

Figure 8.1.1: Representation of a Material Wave



The medium is in *equilibrium* when it is not being disturbed. If the medium above represents a segment of a rope, then it would lie flat in equilibrium. One can disturb this medium by shaking one end of the rope. In the instant of the wave depicted in Figure 8.1.1, section 2 is *displaced* from equilibrium as it pulls on sections 1 and 3. The distance from equilibrium is known as the *displacement* of the medium as the waves passes through it. By Newton's third law, section 2 is pulled down by sections 1 and 3. This is shown by the opposite arrows, $F_{1 \text{ on } 2} = -F_{2 \text{ on } 1}$ and $F_{3 \text{ on } 2} = -F_{2 \text{ on } 3}$, in the force diagrams above. Thus, section 2 will accelerate downward, back toward equilibrium. This will cause section section 3 to accelerate upward, so a little time later section 3 is displaced like section 2 was. In this way, the disturbance has traveled from section 2 to section 3 without the individual pieces of medium traveling along with it.

Alert

When we describe waves we are describing some kind of motion. We will often speak of the waves moving in space to the right, to the left, or outward from the source. This motion in space, does not refer to the particles which comprise the medium. It is the disturbance or the energy which propagates, defined as the wave. This disturbance occurs due to the interaction between neighboring particles in the medium. These particles oscillate about equilibrium in a wavelike manner, but do not physically travel in space along the wave. In other words, material waves provide a mechanism for transferring energy over considerable distances, without the transport of the medium itself.

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 pairwise interactions you learned about in 7A. Individual atoms have a preferred equilibrium separation distance, and resist being pushed or pulled to maintain that preferred distance. Stretching or compressing the medium causes the atoms to exert forces on their

neighbors and to resist forces exerted on them, known as restoring forces. We have simply clumped atoms together into three sections for convenience, but this same discussion applies to individual atoms. The restoring forces responsible for wave behavior are typically strongest in solid materials and negligible in most gaseous materials.

The concepts we introduce when discussing material waves are no different from the concepts that we have already introduced when discussing forces, motion, and atoms. We place such an emphasis on *waves* as a separate phenomenon because dealing with the form of a wave is often simpler than trying to visualize force diagrams for many different sections of a medium. Eventually we will encounter non-material waves, and this 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 (called a *periodic wave*), or the oscillations may exist for only a limited time at each point (a *wave pulse*). For example, if you shake one end of a rope depicted in Figure 8.1.1 only one, you will send a wave pulse down the rope. On the other hand, if you continue to shake the rope in a consistent and periodic manner, then you will generate a periodic wave on the rope.

Waves in a medium are started by outside forces that act on some of the particles in the medium to start them oscillating. The object which generates these external forces is known as the *source* of the wave. External forces are not required to keep a wave going once it has been started. Consider a water wave on the surface of a pond; a rock dropped into the pond is the outside disturbance that starts wave motion. Once started, the ripples expand outward on their own. We don't have to continue dropping rocks into the pond to keep the ripples moving.

Formal Definition of Waves

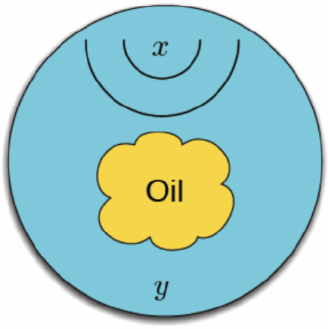
It is difficult to come up with a general definition for a "wave". For the moment we will content ourselves by writing down the following definition of a material wave:

> A *material wave* is the large movement of a *disturbance* in a medium, whereas the *particles* that make up the medium *oscillate* about a fixed *equilibrium position*.

Let's try again to understand what exactly is moving when describing a wave through an example. Consider ripples expanding in a pond. While there is some movement of the individual water molecules, they merely bob up and down, and they do not travel. It is the ripples that move significantly. The movement of the ripples across the surface of the water is what we mean by a wave. It is not the movement of the individual water molecules, or the disturbance of the surface of water from equilibrium, but rather the movement of the disturbance to different locations in the medium that we call the wave.

Try an experiment at home to clarify this distinction of motion when describing a wave. Take a bowl of water, and drop a small amount of olive oil on the surface, so that you have a set-up like picture below. If you oscillate your hand gently at the location x, do you get ripples at the location y? What about the oil? Does it move to the location y?

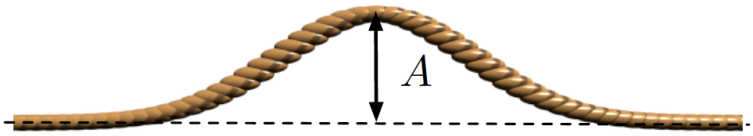
Figure 8.1.2: Wave Motion Experiment



Pulse Wave

A pulse is a disturbance with a finite length. For example, if you shook the end of a rope only once you would produce a pulse wave as shown below.

Figure 8.1.3: Wave Pulse



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 called the *amplitude*, A. We define A to be always positive even if the displacement is below rather than above the equilibrium dashed line shown in the figure above. The amplitude depends on the source, or the amount of energy transferred to the medium by an external object, such as your hand shacking the rope. Another example of a pulse-type wave is the example of a rock thrown into a pond. In this example the rock produces many ripples, but they are a pulse wave because there are a finite amount of ripples with a well-defined beginning and end.

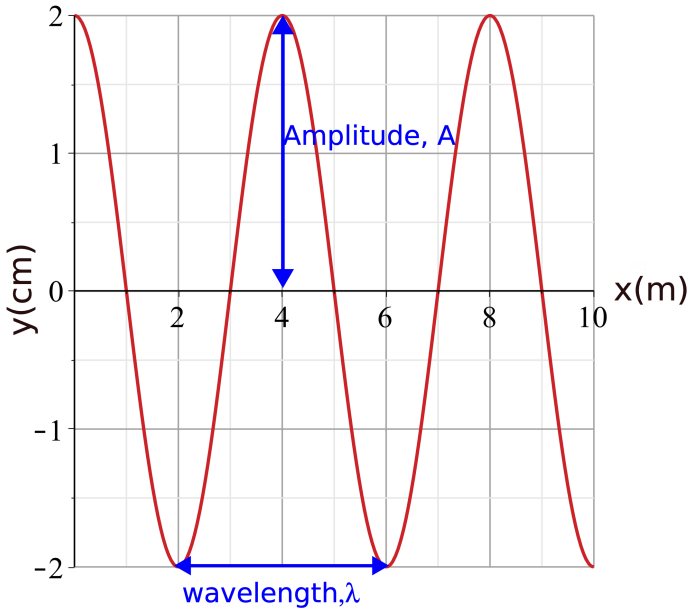
A Periodic Wave

Repeating waves can have many different shapes. One of the simplest to work with looks like a sine or a cosine function. Such waves are called *harmonic* or *sinusoidal* waves. They are generated by oscillators moving in simple harmonic motion, like the spring-mass system you studied in 7A. In other words, a harmonic wave can be modeled with motion of each particle in the medium described by spring-mass oscillator. If you hold one end of a rope and jiggle it up and down in simple harmonic motion, moving your hand in a periodic repeating manner, you will generate harmonic waves. If you were to take a picture of the wave, it might look like the figure below.

Figure 8.1.4: Periodic Wave

When you "take a picture" of a wave you capture its shape at some instance of time, which we call the *snapshot* of the wave. This can be represented graphically as shown below. Here x is the distance along the rope, and the *displacement*, y, represents how far the rope is displaced from equilibrium.

Figure 8.1.5: Graphical Representation of a Harmonic Wave at a Fixed Time ("Snapshot")

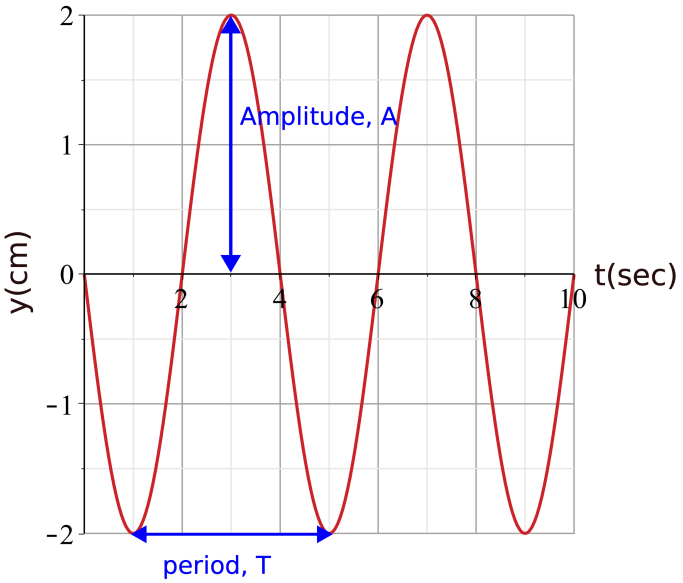


Like the wave pulse, a repeating wave has a well-defined equilibrium position and an amplitude, A. In Figure 8.1.5 the equilibrium position is at y=0 cm, and the amplitude is 2 cm. Compared to a wave pulse, a repeating wave has two new parameters. The first is the *wavelength*, λ, which tells us the shortest distance (along the direction of wave motion) between identical parts of the wave. In other words, the wavelength represents the length of the spatial cycle of the wave as marked in Figure 8.1.5 above. In this figure $\lambda = 4$ m.

The second new parameter is the *period* ,T , the time it takes for the wave to look exactly the same. In other words, the wavelength tells us how the wave repeats in *space*, while the period tells us how the wave repeats in *time*. One period is the time it takes for the wave to move a distance of one wavelength, since it will look the same after one cycle. Oscillators in the medium go through one cycle during the time of one period, such as a spring-mass returning to its starting position. Pulse-type waves, which do not repeat, do not have periods or wavelengths. The wave period is determined by how often the source is disturbing the medium. If you were generating a harmonic wave on a rope, you can determine the period by measuring how long it takes for your hand to move up and back down.

Since the graph in Figure 8.1.5 shows the wave for a fixed time, so it gives us no information about the period which is time-dependent. To give a representation of the temporal characteristic of the wave, we need a new graph. Since there are two variables which describe the wave, space and time, in order to graph the wave changing with time we now need to represent it at a fixed position. If we were to paint a red dot on the rope at some fixed x value, and then plot the position of only that dot against time, we would find how that particular point moves in simple harmonic motion (see Figure 8.1.7 below). Thus, when representing the motion of the wave as a function of time, we are showing the harmonic oscillation of one specific particle in the medium which is at some fixed location in space. It takes the same amount of time for the "dot" to return to an initial position as it does for the whole wave to return to an initial configuration, as you can see in Figure 8.1.7. An example of a displacement versus time graph is shown below.

Figure 8.1.6: Graphical Representation of a Harmonic Wave at a Fixed Position



The period representing one temporal cycle is 4 seconds in the graph above. Assuming that figures 8.1.5 and 8.1.6 represent the same wave, the information they provide must be consistent. For example, both graphs show the same amplitude of 2 cm. The maximum displacement, known as the *crest* of the wave, is at 2 cm, the minimum displacement, the *trough* of the wave, is at -2 cm, and the midpoint between the crest and the trough is the equilibrium position, here at y=0 cm. In Figure 8.1.5 we see that the particle at position x=0 m is at a crest, position x=1 m is at equilibrium, and x=2 m is at a trough. Let us assume that Figure 8.1.5 is a snapshot at t=0 sec. Since Figure 8.1.6 shows t=0 sec at equilibrium, that means that the position it represents has to be at equilibrium in Figure 8.1.5. Thus, possible positions for the time-dependent plot in Figure 8.1.6 could be x=1, 3, 5, etc, where the displacement is at equilibrium in Figure 8.1.5.

The period is the amount of time for one cycle. Another useful measure of the periodicity of a wave is the number of cycles that fit in some unit of time, known as the *frequency*. Frequency is the reciprocal of the period:

$$T = \frac{1}{f} \tag{8.1.1}$$

Typically, period is measured in seconds and frequency is measured in units called *Hertz, Hz* (1 Hz = 1 s^{-1}). The period is the time between the arrival of adjacent crests of the wave, while the frequency is the number of crests that pass by per second. This means the red dot in Figure 8.1.7 completes an f number of up-and-down cycles every second. Equivalently, if you focus on a specific position in space, wave will return to the same configuration an f number of times every second.

Dimensionality

The examples seen above are called *one-dimensional waves* because the wave only travels in one direction. We arbitrarily called that direction the x-axis. If a wave spreads out on a surface instead, then it is a *two-dimensional wave*. For example, ripples on the surface of a pond represent a two-dimensional water wave. A wave that spreads outward in all directions is a *three-dimensional wave*. Examples of three-dimensional waves are typical sound and light waves.

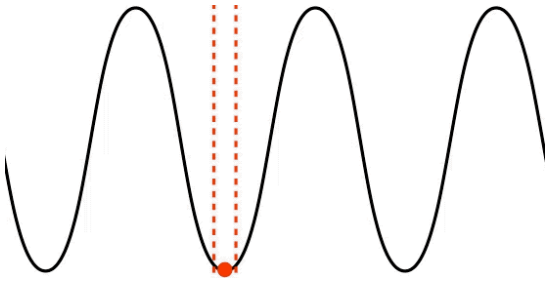
An important distinction between these waves is that the amplitude A of the waves is only constant for one-dimensional waves. Two-dimensional and three-dimensional waves have amplitudes that depend on the distance from the source of the disturbance. Ripples in a pond become smaller as they spread out, and the brightness (amplitude of light) and loudness (amplitude of sound) of light and sound waves, respectively, decreases with distance away from the sources. This is a consequence of conservation of energy, as a wave propagates outward in multiple dimensions, the energy it carries must be spread over a region of increasing size.

Polarization

Material waves and electromagnetic waves have a characteristic called *polarization*. The polarization relates the direction of wave displacement to the direction of wave motion. For material waves, we are going to define two types of polarization:

- *Transverse Polarization*: A material wave is *transverse* if the displacement from equilibrium is perpendicular to the direction the wave is traveling. Below is a transverse wave since the wave is traveling to the right, while the oscillations in the medium are vertical, or perpendicular to the motion. The "red dot" represents one oscillator moving up and down periodically as the wave propagates to the right through the medium. An example of a transverse wave, is a wave generated on a rope or string as in Figure 8.1.4. A surface water wave is another example, since the oscillations of the water particles are in the vertical direction while the wave propagates on the two-dimensional plane of the water surface.

Figure 8.1.7: Transverse Wave



- *Longitudinal Polarization*: A material wave is *longitudinal* if the medium displacement from equilibrium is in the same direction that the wave is traveling. In most examples of longitudinal waves that we explore, this displacement occurs as periodic *compressions* (region of more dense medium) and *rarefactions* (regions of less dense medium) of the material. The figure below shows an example of a longitudinal wave on a spring (or slinky). The most common example of longitudinal waves are sound waves, which we will discuss in more detail in a later section.

Figure 8.1.7: Longitudinal Wave

Although, transverse waves resemble physically the plots that we drew in figures 8.1.5 and 8.1.6, we represent harmonic longitudinal wave exactly the same way using sinusoidal functions. The displacement on the y-axis in figures 8.1.5 and 8.1.6 does not need to represent the physical depiction of the wave, but rather it shows the distance from equilibrium in the medium. As in the transverse wave, the red dot in the longitudinal waves oscillates sinusoidally. Locations of compressions represent crests while rarefaction represent troughs.

Wave Speed

The *wave speed*, v_{wave}, is the speed at which the disturbance propagates through 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. It is not the speed of the oscillations of the individual particles making up the medium. In fact, from 7A we know that the speed of a harmonic oscillator (spring-mass) is not a constant, but changes with distance from equilibrium. The speed is maximum at equilibrium, and decreases as the oscillator gets further away from equilibrium, reaching zero at the maximum distance from equilibrium.

To a good approximation v_{wave} depends only on properties of the medium, not on wave amplitude or frequency. For large waves, or for waves with extreme frequencies, this approximation breaks down. For now, we simplify our discussion by ignoring dependence of wave speed on amplitude (we do not work with big wave in 7C). We will also ignore dependence of speed on frequency, until we discuss refraction of light waves in a later chapter.

The definition of speed from 7B can be written as:

$$\text{speed} = \frac{\text{distance travelled}}{\text{time spent}} \tag{8.1.2}$$

One period is the shortest amount of time before the wave looks exactly the same. When the wave looks exactly the same again, it has moved a distance of one wavelength by definition. Since it take the wave a time of one period to travel a distance of one wavelength, the wave speed can be written as:

$$v_{\text{wave}} = \frac{\lambda}{T} = \lambda f \tag{8.1.3}$$

The second equality above uses the definition of frequency from Equation 8.1.1. As an example of how the medium determines the wave speed we can look at a material wave on a stretched (characterized by tension) medium, such as a rope or a rubber hose. Both transverse waves and longitudinal waves are possible on a stretched strings or ropes. The speed, v_{wave}, of transverse waves on a stretched string depends on the properties of the string that affect its elasticity and its internal properties. For a string that is thin compared to its length, the relation between the wave speed to the string properties is given by:

$$v_{\text{wave}} = \sqrt{\frac{T}{\mu}} \tag{8.1.4}$$

where T is the tension in the string and μ is its mass density, mass per unit length ($\mu = m/L$). We can make some sense of the formula by considering the diagram in Figure 8.1.1. The tension is (roughly) the force that one piece of string exerts on another. The tighter the string the higher the tension. As we learned, a material wave is a disturbance that propagates by one piece of the medium exerting a force on its neighbors, it is logical that an increase in tension results in an increase in the wave speed. Stronger forces will lead to greater acceleration. When the string is particularly heavy, the forces between pieces of the string result in less acceleration, so it also intuitive that as μ increases the wave speed decreases. The ability to control the wave speed is critical for stringed instruments like a guitar or a violin, which is why they have tuning knobs at one end (to control the tension), and the strings are of different mass (for different values of μ). We will discuss string instruments in more detail when we cover standing waves in later sections.

8.2: Wave Representation

Simple Harmonic Motion

For the rest of the course we will focus on infinite repeating waves of a specific type, *harmonic waves*, whose properties were described in the previous section. We saw how to represent the wave graphically and defined some important wave properties. In this section we want to write down a mathematical expression for one-dimensional harmonic waves and make connections between the mathematical and graphical wave representations.

In order to represent the wave mathematically, let us start by thinking of one oscillator in the medium. We discussed in the previous section that when a wave propagate through a medium, the motion of each particle in the medium can be treated as an oscillator around the equilibrium position of the medium. When the wave is harmonic, these oscillations can be described by a simple harmonic oscillator, similar to the spring-mass system you studied in 7A as depicted in the figure below.

Figure 8.2.1: Simple Harmonic Oscillator

In 7A we focused on describing the energy of the harmonic oscillator: the potential and kinetic energy relative to the equilibrium position. Now we want to focus on the dynamics of the harmonic oscillator, in order to describe how the position of the oscillator changes with time. This result can be obtained by using Newton's second law and the force for a simple harmonic oscillator introduced in 7A in Section 2.5. Hooke's Law describes the restoring force for simple harmonic motion:

$$F_{net} = -ky \tag{8.2.1}$$

where k is the spring-constant, and y is displacement from equilibrium. Combining Hooke's Law with Newton's second law we get:

$$F_{net} = ma = -ky \tag{8.2.2}$$

Using the definition of acceleration and rewriting the above equation such that it is solvable for position as a function of time, we get:

$$m\frac{d^2y}{dt^2} + ky = 0 \tag{8.2.3}$$

From the above equation we can see that y has to be represented by a function that when differentiated twice it returns to the negative of same function. The functions that have this property are sinusoidal, either sine or cosine. As a result the posltion of a harmonic oscillation as a function of time can be represented as:

$$y(t) = A\sin\left(\omega t + \phi_o\right) \tag{8.2.4}$$

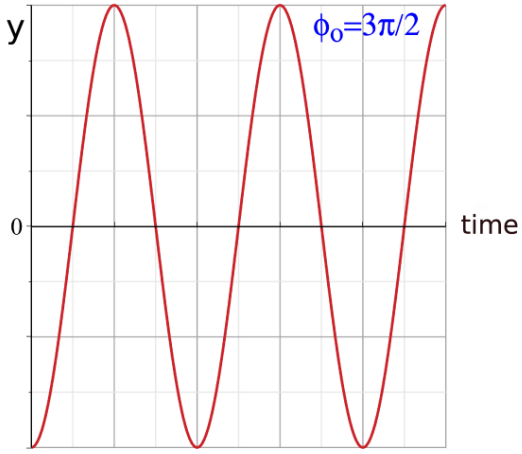
where $\omega = \sqrt{k/m}$ (you will show this result in the Example 8.2.1 below). It is reasonable that harmonic motion which is periodic is represented by a sine function which repeats. The amplitude, A, is the maximum displacement, since the sine function can at most be equal to one. For the spring-mass system the amplitude represents the distance that you initially pulled the spring-mass away from equilibrium before releasing it. We learned from 7A that due to conservation of energy the spring-mass displacement cannot be bigger than the distance from which it is released.

The value ϕ_o inside the sine function is a constant which determines the initial displacement (at t=0 sec) of the system. For example, if you pulled the spring-mass down and released it, such that the displacement is $y = -A$ at $t = 0$ sec, then:

$$y(0) = -A = A\sin(\phi_o) \;\Rightarrow\; \sin(\phi_o) = -1 \;\Rightarrow\; \phi_o = \frac{3\pi}{2} \tag{8.2.5}$$

The sine function in Equation 8.2.4 for the above value of $\phi_o = 3\pi/2$ is plotted below, starting from a maximum displacement below equilibrium.

Figure 8.2.2: Phase Constant Example

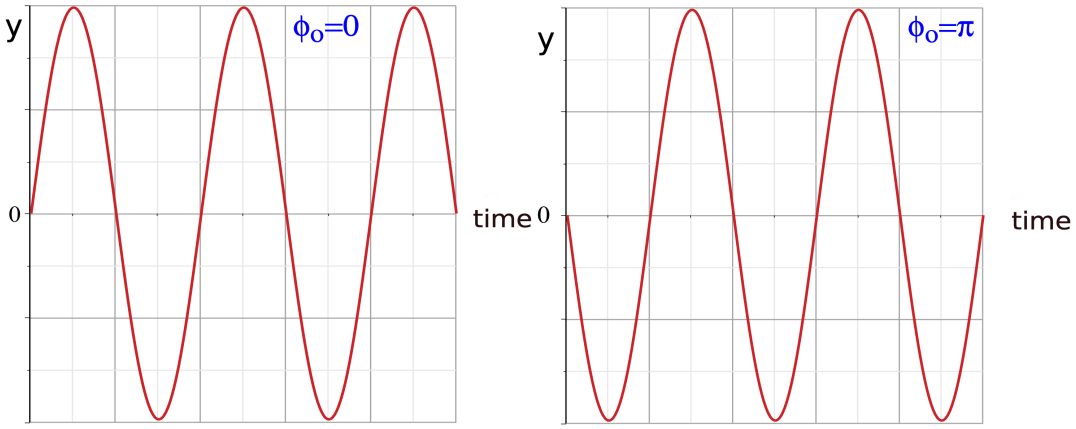


In another example, if you define the initial conditions (t=0s) at the time the spring-mass passes equilibrium (y=0), then:

$$y(0) = 0 = A\sin(\phi_o) \tag{8.2.6}$$

The above scenario is more complicated, since there are two solutions in one cycle of the sine function for this initial scenario, $\phi_o = 0$ or $\phi_o = \pi$. These phase constants correspond to two physical situations that are possible within one cycle when the mass is at equilibrium. If the mass is vertical as in Figure 8.2.1, it could either be moving down or up as it passes equilibrium. In the figure below it can be seen that a starting position of the spring-mass at equilibrium moving up corresponds to $\phi_o = 0$, while the spring-mass moving down at equilibrium at t=0 sec results in $\phi_o = \pi$, the left figure is shifted by half a cycle to obtain the right figure.

Figure 8.2.2: Phase Constant



By definition of one period T introduced in Section 8.1, the function should return to its original position after one cycle. For trigonometric functions a convenient unit for counting cycles is *radians*. Radians of 2π represent one cycle, such that:

$$sin(x + 2\pi) = sin(x) \tag{8.2.7}$$

To assure that the quantity inside the sine function is in units of radians, Equation 8.2.4 can be written as:

$$x(t) = A\sin\left(\frac{2\pi}{T}t + \phi_o\right) \tag{8.2.8}$$

where ω Equation 8.2.4 is written in terms of $2\pi/T$. To check that this expression is consistent with the definition of one period, the function should be identical when one period of time passes, $t \to t + T$:

$$y(t+T) = A\sin\left(\frac{2\pi}{T}(t+T) + \phi_o\right) = A\sin\left(\frac{2\pi}{T}t + 2\pi + \phi_o\right) = A\sin\left(\frac{2\pi}{T}t + \phi_o\right) = y(t) \tag{8.2.9}$$

In the above equation we used the property of the sine function given in Equation 8.2.7.

Example 8.2.1

Show that that $\omega = \sqrt{k/m}$ using Equation 8.2.3 and Equations 8.2.4.

Solution

Starting with Equation 8.2.4:

$$y(t) = A\sin\left(\omega t + \phi_o\right)$$

Differentiating it once:

$$\frac{dy}{dt} = A\omega\cos(\omega t + \phi_o)$$

Now differentiating it again:

$$\frac{d^2y}{dt^2} = -A\omega^2\sin(\omega t + \phi_o)$$

Now plugging the above results into Equation 8.2.3:

$$-A\omega^2 m\sin(\omega t + \phi_o) + kA\sin(\omega t + \phi_o) = 0$$

Cancelling the sine terms and the amplitude:

$$-\omega^2 m + k = 0$$

Solving for ω:

$$\omega = \sqrt{\frac{k}{m}}$$

Harmonic Wave Equation

The discussion above which led to Equation 8.2.8 does not capture the full wave phenomena since it only describes the motion of one oscillator at fixed position in the medium as a function of time. That "fixed position" refers to the location in the medium about which the particle oscillates. To describe the wave fully, we also need to incorporate the wave motion in space throughout the medium. In other words, we want to write down a mathematical expression describes the displacement of all the oscillators in the medium at all times. A one-dimensional harmonic wave can be expressed mathematically as:

$$y(x,t) - y_o = A\sin\left(\frac{2\pi}{T}t \pm \frac{2\pi}{\lambda}x + \phi_o\right) \tag{8.2.10}$$

The displacement away from equilibrium at position x and time t is given by the left-hand side, $y(x,t) - y_o$, where y_o is equilibrium position, and $y(x,t)$ is the oscillator's position relative to some chosen zero. In most cases it is convenient to set $y_o = 0$, since this will result in a symmetric displacement about the x-axis with the maximum displacement at A and the minimum displacement at $-A$. But when the equilibrium position is not set to zero, then the entire sine function shifts vertically either up or down by the chosen value of y_o.

Equation 8.2.8 demonstrated that for a single harmonic oscillator the sine function has identical form after a time of one period. The second term in Equation 8.2.10, which describes the spatial periodicity of the wave, resembles the first term very closely. Using the same method as in Equation 8.2.9, one can show that the wave equation is identical after the value x increases by any multiple of wavelength.

The *fixed phase constant* ϕ_o is different from the one in Equation 8.2.8, since it describes what the wave looks like both at the initial time, $t = 0$, and at a location defined as zero, the origin $x = 0$. For example, assume that there is a crest at $x = 0$ and $t = 0$. If we set the equilibrium at zero $(y_o = 0)$, then the initial displacement of the wave is equal to A, since it is a crest:

$$y(0,0) = A = A\sin(\phi_o) \;\Rightarrow\; \sin(\phi_o) = 1 \;\Rightarrow\; \phi_o = \frac{\pi}{2} \tag{8.2.11}$$

The symbol $\pm$ asks us to choose the sign of either $+$ or $-$ depending on the direction that the wave travels. How do we determine the appropriate sign when the wave is moving to the right (values of x are increasing) or when the wave is moving to the left (values of x are decreasing)? The quantity $y(x,t)$ gives the displacement of a particular part of the wave as it travels through space and time. So if you are watching a particular part of the wave, the displacement should stay fixed for all values of position and time. Let's say you start watching a

crest at $t = 0$ of a wave that starts at $x = 0$ and moves to the right. If you watch the crest for a time span of a quarter of a period, $t = T/4$, then this crest must be at a position quarter of cycle in the direction of increasing x, $x = \lambda/4$. Using the result in Equation 8.2.11 for the phase constant when the initial conditions of the wave is a crest, the displacement at a time $t = T/4$ is:

$$y\left(\frac{\lambda}{4}, \frac{T}{4}\right) = A\sin\left(\frac{2\pi}{T}\cdot\frac{T}{4} \pm \frac{2\pi}{\lambda}\cdot\frac{\lambda}{4} + \frac{\pi}{2}\right) = A\sin\left(\frac{\pi}{2} \pm \frac{\pi}{2} + \frac{\pi}{2}\right) = \begin{cases} + \Rightarrow & A\sin\left(\frac{3\pi}{2}\right) = -A \\ - \Rightarrow & A\sin\left(\frac{\pi}{2}\right) = A \end{cases} \tag{8.2.12}$$

Since you were tracking a crest, the correct sign to choose is negative since it resulted in a displacement equal to A, according to the above equation. If the wave was moving to the left instead, then its position a quarter period later would be at $x = -\lambda/4$, and following the same steps are outline above you would find that a positive sign in front of the spatial term would give the correct result. To summarize, we choose a "+" sign for left-moving waves and a "−" sign for right-moving waves when writing down the wave equation given in Equation 8.2.10.

There is another very useful quantity which we will often use to describe wave properties. The *total phase, Φ,* (not to be confused with the fixed phase constant ϕ_o) is defined as the quantity inside the sine function of the wave equation:

$$\Phi(x,t) \equiv 2\pi\frac{t}{T} \pm 2\pi\frac{x}{\lambda} + \phi_o \tag{8.2.13}$$

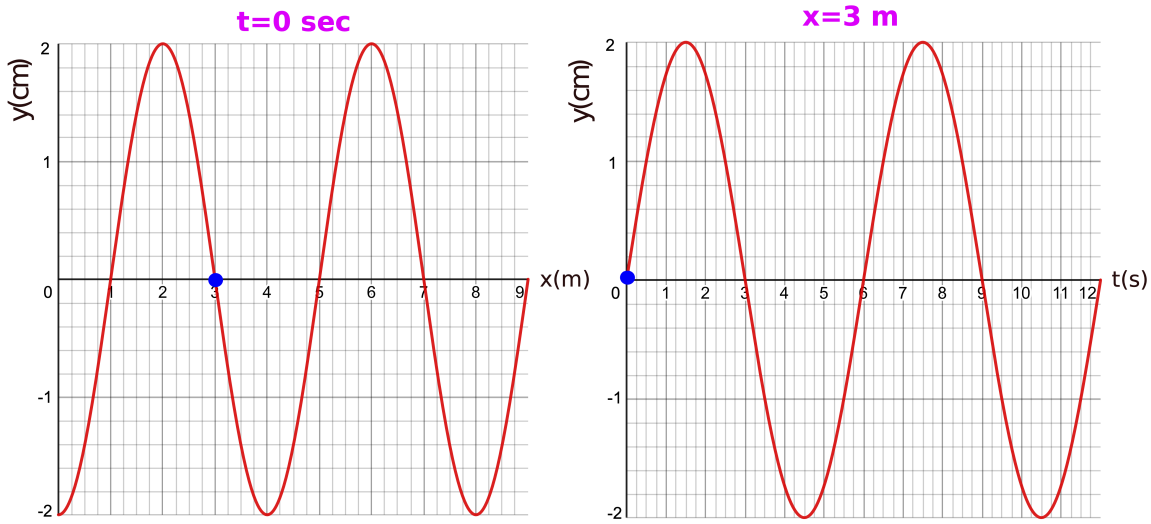
The total phase identifies a particular point on a wave. When we imagine ourselves riding the wave, or when we watch a wave crest or trough travel, we are really following a point of *constant total phase.* Going back to the direction discussion, since the total phase of any part of the wave must remain constant, if the wave is moving to the right (time and position are both increasing), the sign in front of the spatial term must be negative to keep Φ constant. On the contrary, if the wave is moving to the left (time is increasing while position is decreasing), then the sign must be positive to keep the total phase fixed.

Since the wave equation has many terms, it is worthwhile noting the difference between *variables* and *parameters,* and the dependence of parameters on the source of the wave or the medium through which it travels:

- The parameters $A, T, \lambda, \phi_o, y_o$ and the choice of $+$ or $-$ are constants for any given harmonic wave. They describe the wave and its behavior. The period and the amplitude are determined by the source of the wave. The phase constant, the equilibrium position, and the direction is determined by the choice of origin and coordinate system, but remain fixed once they are established. The wave speed is determined by the medium. The wavelength depend on both the period and the wave speed.
- The variables (for any x and any t) describe the all the space and time in which the wave exists. These variables determine the displacement of the medium at that particular time and location.
- We can ask about the displacement at different locations and times by changing variables, but parameters for a wave are fixed values.

Now that we have an equation to fully describe one-dimensional wave phenomena, we want to see if it is possible to represent it graphically. Typically, you can plot some behavior by taking the equation which it describes and plotting it as a function of the variable. But for the wave equation there are two variables, space and time. So, you can fix a particular time and make a plot of the wave as a function of position, which we called the snapshot plot in the previous section, which does not contain any temporal information about the wave. Or you can look at the medium at a specific location (fix the position) and plot how that part of the medium oscillate with time, but then you loose the spatial characteristics of the wave. Thus, it is clear that need more than one plot to fully describe wave phenomena. But are only two plots enough to describe the wave all all positions and all instances of time? Let us start by analyzing the two plots shown below.

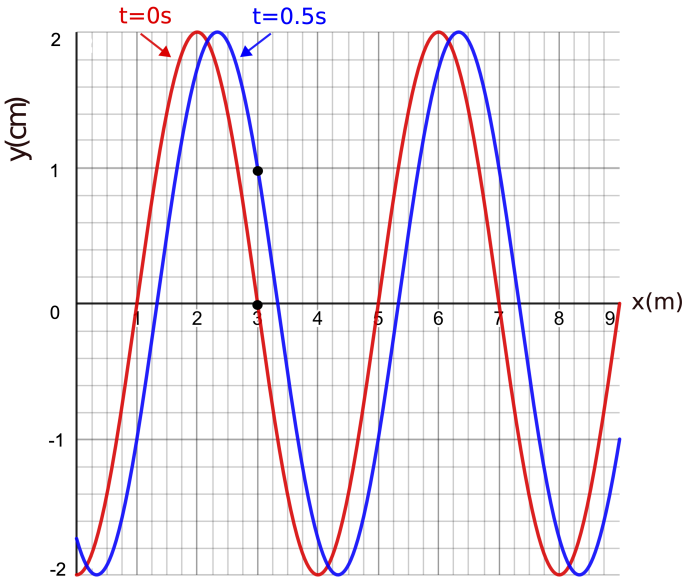
Figure 8.2.3: Graphical Representation of a Wave



The plot on the left is a snapshot plot of the wave taken at $t = 0$ sec. You can read from the plot the amplitude of the wave, $A = 2$ cm, and the wavelength, $\lambda = 4$ m. The plot on the right show the oscillation of the medium at $x = 3$ m as a function of time. It shows the same amplitude which has to be consistent between the two plots. The plot also provides information about the period, $T = 6$ s. From both plots we can see that the equilibrium is set at zero, $y_o = 0$.

To fully write down the wave equation, we still need to figure out the wave direction and the fixed phase constant. Let us think about direction first. We are given a snapshot (a paused "movie" of the moving wave) a $t = 0$ sec. We want to figure out whether we have enough information from the two plot which will allow us to determine if the wave will shift to the right or the left as some later time, when we un-pause the "movie". The only additional information we do have is the displacement at $x = 3$ m position will increase after $t = 0$ sec, or a crest is moving toward the $x = 3$ m location. The two (blue) dots shown in Figure 8.2.3 represent the same position and time ($t = 0$ s and $x = 3$ m). In other words, the two dots are the common point between the two plots, which happens to be an equilibrium displacement. Since the right plot tells us that the displacement at $x = 3$ m will increase after $t = 0$ s, shifting the snapshot plot in the correct direction of motion should give the same result. Since we want a crest to approach the $x = 3$ m position, then the snapshot plot needs to be shifted to the right. Figure 8.2.4 below stresses this point by plotting the shifted snapshot at a later time of 0.5 seconds.

Figure 8.2.4: Shifting the Snapshot Plot



The time plot in Figure 8.2.3 shows that at $t = 0.5$ s the displacement at x=3m is going to be around 1 cm. The time of 0.5 seconds is 1/12 of one cycle since one cycle is 6 seconds. This corresponds to the wave moving by 1/12 of one wavelength or 1/3 m, which can be approximately seen by the difference in the two overlapping plots in the figure above. After the shift, the displacement at x=3m went up to 1 cm, as expected. If you were to shift the plot to the left instead, then the displacement would decrease, which is not consistent with the time-dependent plot. Therefore, we conclude that this wave is moving to the right.

Lastly, let's see if it is possible to obtain the fixed phase constant from these two plots. We defined ϕ_o as the displacement of the wave at $t = 0$ s and $x = 0$ m. From the snapshot in Figure 8.2.3 you can see that this value corresponds to a trough. Plugging this into Equation 8.2.10 we get:

$$y(0,0) = -2cm = (2cm)\sin(\phi_o) \Rightarrow \sin(\phi_o) = -1 \Rightarrow \phi_o = \frac{3\pi}{2} \tag{8.2.14}$$

This was of phase constant was simply to obtain. However, not all values of fixed phase constant could be easily read from the plot. For example, a fixed phase constant of $\pi/4$ corresponds to displacement of 0.7071 (if amplitude is one), which would be hard to interpret from the plot. But there is no need to only use the values at the origin and initial time. You can use any point on the plot and just solve for the phase constant. For example, from the snapshot plot in Figure 8.2.3, we can see that at $x = 2$ m, the displacement is a crest. Since the snapshot plot is at $t = 0$ sec, plugging into the wave equation, and using the parameters which were already determined we get:

$$y(x = 2m, t = 0s) = 2cm = (2cm)\sin\left(\frac{2\pi}{6s}(0) - \frac{2\pi}{4m}(2m) + \phi_o\right) = (2cm)\sin(-\pi + \phi_o) \Rightarrow \sin(-\pi + \phi_o) = 1 \tag{8.2.15}$$

The total phase has to be equal to $\pi/2$ for a crest, since $\sin(\pi/2) = 1$. Therefore, solving for the phase constant we get:

$$-\pi + \phi_o = \frac{\pi}{2} \Rightarrow \phi_o = \frac{3\pi}{2} \tag{8.2.16}$$

which gives a result identical to the one found in Equation 8.2.14. You can choose any other location from the two plots, and you will get the same result for the fixed phase constant. In general, it is always helpful to choose either a crest or a trough to solve for ϕ_o, since there is only one value for the total phase, $\pi/2$ for crests and $3\pi/2$ for troughs. You can also use an equilibrium value but there are two solution for the total phase, 0 or π, so you would need to decide which one is the correct result, depending on how the function is changing after the equilibrium location that you chose. Any other value except for the equilibrium, a crest, or a trough may be difficult to read from the plot.

There are infinite number of solutions for the fixed phase constant, since you can also add or subtract any multiple of 2π to the phase constant and obtain the same result for the wave equation. For example, another possible solution for the phase constant in Equation 8.2.14 is $-\pi/2$ or $7\pi/2$ or even $23\pi/2$. In fact, if you were to choose another crest or trough to work with, you might obtain a number different from $3\pi/2$. In general, for simplicity it is best to keep the value of phase constant between $-\pi$ and π.

Putting everything together the wave equation for the wave depicted in Figure 8.2.3 is:

$$y(x,t) = (2cm)\sin\left(\frac{2\pi}{6s}t - \frac{2\pi}{4m}x + \frac{3\pi}{2}\right) \tag{8.2.17}$$

Note, it is important to write units for the parameters such as amplitude, wavelength, and period to make sure that correct units are used when plugging values into x and t and solving for $y(x,t)$.

Example 8.2.2

Two buoys are oscillating up and down as ripples in a pond continuously pass through them. At t=6 sec you observe that the buoy which is at position x=4 m is at equilibrium, and the second buoy at position x=30 m is at a trough. You also count three crests and three troughs between the two buoys at t=6 sec. The vertical distance between the two buoys at t=6 sec is 0.2 m. The frequency of the wave is 0.5 Hz.

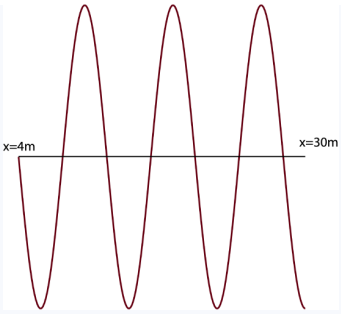
a) Calculate the speed of the wave.

b) At t=6.5 sec the buoy which is at x=4 m is found at a crest. Is the wave moving to the right or the left?

c) Write down the full wave equation for this wave.

Solution

a) To find the speed we are missing direct information about the wavelength of the wave. It is helpful to visualize this problem by making a picture of the two buoys at t=6 sec. The x=4m buoy is at equilibrium and the x=30m buoy is at a trough, and there are 3 crests and troughs in between them.



From the picture we see that there are three and one quarter cycles between these points which gives direct information about the wavelength:

$$3.25\lambda = (30-4)m$$

$$\lambda = \frac{26m}{3.25} = 8m$$

Using the given frequency, we can calculate the speed of the wave:

$$v = \lambda f = (8m)(0.5Hz) = 4m/s$$

b) From 6 seconds to 6.5 seconds, 0.5 seconds passed. The period is $T = 1/f = 2s$, so 0.5 sec is one quarter of a cycle. There is a crest one quarter cycle to the left of the buoy at x=4m (see picture above), so the wave is moving to the right.

c) Since the vertical distance between the two buoys is the distance between equilibrium and a minimum displacement, the amplitude is 0.2 m. The last parameter left to determine is the phase constant. It is simplest to choose a crest or a trough at a specific time and location. From the picture above, let us use the crest at 10m, which is a three quarters of a wavelength to the right of the buoy at x=4m. Plugging into the wave equation:

$$y(x = 10m, t = 6s) = (0.2m)\sin\left(\frac{2\pi}{2s}\cdot 6s - \frac{2\pi}{8m}\cdot 10m + \phi_o\right) = 0.2m$$

Which results in the total phase being equal to $\pi/2$:

$$6\pi - \frac{5\pi}{2} + \phi_o = \frac{\pi}{2} \Rightarrow \phi_o = -3\pi$$

Note, although mathematically the result for the fixed phase constant came to -3π, this is equivalent to π due to the sine function property that changing the phase by any multiple of 2π does not change the sine function. And it is always more presentable to express the fixed phase constant with a value between $-\pi$ and π, although it is not technically wrong to leave it outside this range. So we choose:

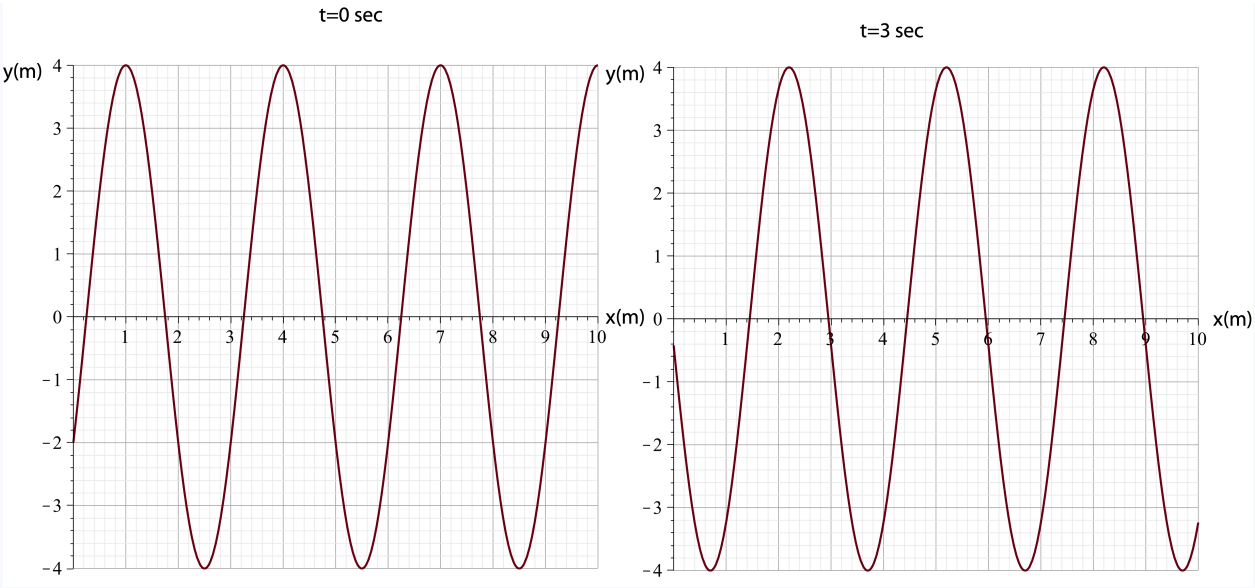
$$\phi_o = \pi$$

Combining all the results, the wave equation for this wave is:

$$y(x,t) = (0.2m)\sin\left(\frac{2\pi}{2s}t - \frac{2\pi}{8m}x + \pi\right)$$

Example 8.2.3

You are enjoying spring break at the ocean but decide to prepare for 7C by analyzing ocean waves. You take pictures of the waves 3 seconds apart and make two plots of wave displacement over 10 meters. You also note that the crest in at x=7 m and t=0 sec is the same crest as the one at x=5.2 m and t=3 sec. Write down the wave equation for this wave.



Solution

The wave moves to the left since the crest moves in the negative x-direction (from 7m to 5.2m) as time moves forward. The crest moves 1.8 meters in 3 second, so the wave speed is:

$$v = \frac{d}{t} = \frac{1.8m}{3s} = 0.6m/s$$

The wavelength is 3 meters from the plots. Solving for period:

$$T = \frac{\lambda}{v} = \frac{3m}{0.6m/s} = 5s$$

Let us use the crest at t=0sec and x=1m to find the phase constant:

$$y(x = 1m, t = 0s) = (4cm)\sin\left(\frac{2\pi}{5s}\cdot 0 + \frac{2\pi}{3m}\cdot 1m + \phi_o\right)$$

Since it is a crest:

$$\frac{2\pi}{3} + \phi_o = \frac{\pi}{2}$$

Resulting in:

$$\phi_o = -\frac{\pi}{6}$$

Combining all the results, the wave equation for this wave is:

$$y(x,t) = (4cm)\sin\left(\frac{2\pi}{5s}t + \frac{2\pi}{3m}x - \frac{\pi}{6}\right)$$

8.3: Multi-Dimensional Waves

Sound Waves

Sound is a wave in the sense that we have already defined. However, sound waves in air behave differently from the material waves we have described so far. As we discussed in Physics 7A the bonds between particles in the air are virtually non-existent, so the particles in the medium are non-interacting and do not exert forces on each other. The oscillations in the medium can no longer be described in terms of restoring forces modeled by Hooke's Law. Instead, we can explain how sound wave propagate through air in terms of variations in pressure. In other words, a sound wave can be described as a *pressure wave*. The pressure variations occur in the same direction as wave propagation, so sound waves are longitudinal.

Although, individual air particles do not oscillate about some equilibrium position, as a sound wave travels through air the oscillation occur in the density or the pressure of the air. We learned in Physics 7A that air molecules move around rapidly in random motion, but there are well-defined averages for density and pressure. Sound waves in air are the oscillation of the average value of particle density (and therefore pressure) over distance scales much larger than the mean distance between particles. Thus we can model sound wave considering only the pressure or density of the air.

While this is slightly different in form from the material waves, 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.

Choosing pressure to describe sound waves, the sound wave equation becomes:

$$P(x,t) - P_{atm} = P_o \sin\left(\frac{2\pi}{T}t \pm \frac{2\pi}{\lambda}x + \phi_o\right) \tag{8.3.1}$$

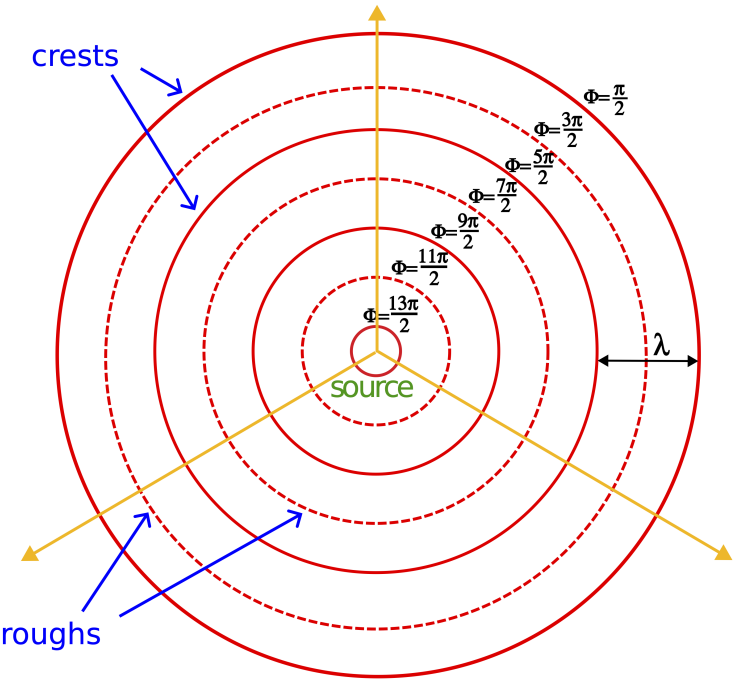
where $P(x,t)$ is the absolute pressure of the air at a given position x and at a time t. P_{atm} is the atmospheric pressure (i.e. the equilibrium pressure). P_o is the amplitude of the pressure fluctuation (gauge pressure) from equilibrium, such that when the displacement is at a crest, the pressure is at a maximum of $P_{atm} + P_o$, and when it's at a trough, the pressure is at a minimum of $P_{atm} - P_o$.

Note the similarities between this equation in terms of pressure and Equation ??? in terms of individual particle displacement. We can therefore ascribe familiar parameters to sound waves, such as wavelength, period, and wave speed. We can also use the same techniques from the previous section to plot the pressure against time at constant position, or pressure against position at constant time.

Wavefronts and Rays

The wave equation we introduced in the previous section describes one-dimensional waves. However, in most cases waves propagate in either two-dimensions (such as the surface of water) or in three-dimensions (such as sound or light). Although, we will not introduce the mathematical representation of higher dimensional waves in this course, there are other representations for multi-dimension waves which are useful. The figure below shows a drawing of two-dimensional *wavefronts*, concentric circles are crests (solid lines) and troughs (dashed lines) emanating from the source located at the center. The waveforns spread as they move away from the source, like ripples in a pond.

Figure 8.3.1: Wavefronts and Rays



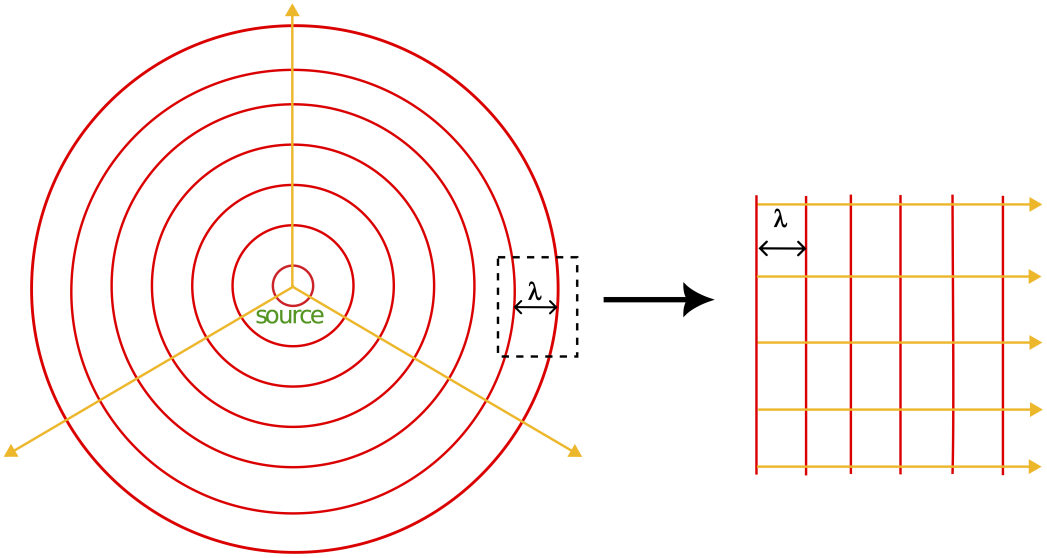
This is a snapshot representation of the location of crests and troughs at a specific time. At a later time, each wavefront would be found further from the source and more wavefronts would be shown as new crests are troughs are generated by the source. Each wavefront provides information about the relative times that they were generated Since the wavefronts move outward, the wavefront which is furthest from the source was the first one that was generated, and the one closest to the source is the last one to be emitted. Thus, time "increases inward" from the furthest to the closest wavefront. If we define $x = 0$ m to be the location of the source, and $t = 0$ s is the time the first crest was generated, then the total phase of the furthest crest is $\pi/2$. The furthest trough was generated half a cycle later, at $t = T/2$ s, so the total phase is bigger than that of the first crest's by π, resulting in a total phase of $3\pi/2$. The phase of the next crest is a full cycle larger, or 2π bigger than the phase of the first crest, and so on. The wavefront representation stresses that the total phase is a characteristic of a specific part of the wave.

The wavefront representation also contains information about the spacial periodicity of the wave, since distance between any neighboring crests or troughs is one wavelength. However, since this is a snapshot, information about the temporal periodicity, period or frequency, is lost in this representation. Also, since this is a "top view" of the wave propagating in two-dimensions, information about amplitude cannot be obtain from this picture. In fact, as we will see shortly, amplitude is no longer a constant in higher dimensional waves. Although complete information about the wave in lost in this representation, it is often an informative tool for understanding certain wave phenomena, such as interference introduced in later sections. In three-dimensions the concentric circles in Figure 8.3.1 would become concentric spherical shells.

The arrows drawn in the figure provide yet another useful wave representation. The arrows only preserve the information about wave direction. They are always drawn perpendicular to the wavefronts and point in the direction of wave motion. This representation will be especially useful when we discuss optics as an application of various properties of light. Various optical phenomena that we will study mostly arise due to the direction of the motion of rays of light. Thus, the most useful representation for that unit will be the *ray representation*.

Most waves in nature are multi-dimensional. The only truly one-dimensional wave are those confined to a string or rope. However, the one-dimensional wave representation is often a good approximation to multi-dimensional waves far away from the source. Imagine an observer standing far away from a wave source and can only detect a small area of the wavefronts. This is represented by the dashed rectangle in the figure below. To this observer the individual wavefronts look almost parallel. The further the observer is from the source, the smaller portion of the wavefronts they can detect, the more parallel the wavefronts will appear. If the wavefronts are parallel that means the wave is moving in one direction perpendicular to the wavefronts. This is illustrated on the right "zoomed-in" picture of the wavefronts far from the source. The rays which are perpendicular to the wavefronts are all parallel, implying one-dimensional motion of the waves.

Figure 8.3.2: Plane Waves



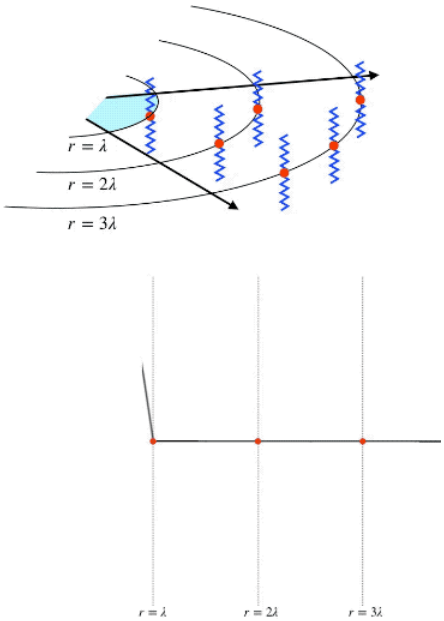
This approximation of multi-dimensional waves distant from the source is known as *plane waves*. Thus, the one-dimensional wave equation is often used to represent plane waves which originate from a source generating multi-dimensional waves. The distance between two neighboring crests is preserved to measure one wavelength.

Wave Intensity

In the previous section we mentioned that amplitude is only preserved in one-dimensional waves. This is assuming that dissipative effects of the medium are negligible. That is, the particles in the medium that oscillate do so without "friction." This means we are assuming that all of the energy in the wave remains within the wave, and none of the energy is converted into thermal energy in the medium. However, even neglecting friction amplitude no longer remains constant for multi-dimensional waves.

Consider now a wave radiating outward from a point source in two dimensions (think of a circular ripple on a pond caused by a pebble). Each position in the medium contains a particle oscillating harmonically (like a mass on a spring), and as the wave propagates outward, the number of oscillating particles increases. The particles in the medium are spaced the same everywhere, so the number of particles encountered by the circular wave is proportional to its circumference, and therefore proportional to its radius. This means that when the radius of the wave front doubles, it is oscillating twice as many particles in the medium. As the wave moves out, there is no energy lost, so when the circle enlarges, the energy is distributed amongst a larger number of oscillators. The energy in each oscillator is determined by its amplitude of oscillation, so for more oscillators to have the same energy as fewer oscillators, their amplitudes must decrease.

Figure 8.3.3: Circular Wave Energy Conservation



Let us see how the energy of the oscillator is related to amplitude. Recall, from 7A the total energy of a spring-mass system is the sum of kinetic and potential energy:

$$E_{\text{tot}} = KE + PE_{sm} = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \tag{8.3.2}$$

Amplitude is defined to be the maximum displacement from equilibrium (x=A), when the speed of the oscillator is zero (turnaround point). This results in the following expression for the total energy:

$$E_{\text{tot}} = \frac{1}{2}kA^2 \tag{8.3.3}$$

Thus, the energy per oscillator is proportional to the *square* of the amplitude. This means that doubling the radius of the circle which doubles the number of oscillators between which this same energy is shared, the square of amplitude of each oscillation reduces by half, and the amplitude by a factor of $\sqrt{2}$. Tripling the radius reduces the amplitude by a factor of $\sqrt{3}$, and so on. In other words, amplitude is proportional to the inverse square-root of the radius:

$$A \propto \frac{1}{\sqrt{r}} \tag{8.3.4}$$

The figure above shows what happens to the amplitude of the wave in cross-section as it goes from a radius of 1 wavelength to 3 wavelengths.

The wave doesn't change its velocity from the inner circle to the outer circle, so the rate at which energy passes through each circle must be the same. What is different about two circles is the *density* of the energy contained in each. For the smaller circle, the energy is distributed over a smaller circumference than for the larger circle, so the energy *density* becomes smaller as the wave propagates outward, even though the total energy is constant. We can define power density in the same manner – by dividing the power of the wave (which is the same for both rings, and everywhere else) by the size of the region through which it is passing. This "power density" is called *intensity*. For our two-dimensional wave, this is the ratio of the power of the wave and the circumference of the circle through which it is passing:

$$I_{2d}(r) = \frac{P}{2\pi r} \tag{8.3.5}$$

Therefore, the intensity of a two-dimensional wave radiating outward from a central point varies in inverse proportion to the distance from the central source. We find that the intensity is proportional to the square of the amplitude:

$$A \propto \frac{1}{\sqrt{r}} \quad \Rightarrow \quad I \propto A^2 \tag{8.3.6}$$

It turns out that the proportionality of intensity and square amplitude was the case for one-dimension as well. For a one-dimensional wave, the energy density does not change, because all of the energy is handed from one oscillator to another neighboring single oscillator. Therefore the power density (intensity) doesn't change, which is consistent with what we already know; the amplitude of a one-dimensional wave remains constant.

Far more common in our studies are three-dimensional waves with central sources (namely sound and light), and the power density in these cases involves dividing by a spherical surface area, rather than a circle. In this case, the intensity of the wave has units of watts per square meter (whereas the intensity of the two-dimensional wave had units of watts per meter), and we have:

$$I_{3d}(r) = \frac{P}{4\pi r^2} \tag{8.3.7}$$

Once again we find the same relationship between intensity and amplitude. The same mechanism is at work: as the wave moves outward from a central point, the number of oscillators on each spherical surface is proportional to the surface area. Doubling the radius of a spherical surface quadruples the surface area, so the number of oscillators grows with the square of the radius. This means that the energy per oscillator drops with the square of the radius, and the amplitude is inversely-proportional to the radius:

$$A \propto \frac{1}{r} \quad \Rightarrow \quad I \propto A^2 \tag{8.3.8}$$

The relation between intensity and amplitude is therefore universal among waves, and one that we will keep in mind in the sections to come.

8.4: Doppler Effect

Stationary Source and Observer

Up to this point when we considered wave properties such as frequency, wavelength, and speed, we assumed that the *source*, which is generating the wave, and the *observer*, who is detecting the wave and measuring its properties, are both *stationary*. The animation below represents this situation where the central (red) dot is the source which flashes every time is emits a crest. The wave representation you see below is of a two-dimensional wave where the circles represent crests emanating out from the source, such as ripples in a pond. The distance between two consecutive circles is the wavelength of the wave. Since the representation below is an amination, you can also determine the period by timing how often the red dot flashes.

Figure 8.4.1: Source and Observer are Stationary

The blue dot to the left of the source is the observer which flashes every time that a crest passes through it. The idea of an observer is especially applicable to sounds waves, which we interpret directly in our daily lives. When the crests exert pressure on our ears, our eardrums vibrate. The amplitude of these vibrations is interpreted by our brains as loudness, and frequency of the vibrations is interpreted as the pitch of the sound.

When both the source and the observer are stationary, the frequency according to both of them will be identical, as will be the wavelength (the measurement of the distance between two consecutive crests), and therefore the speed of the wave. The question we would like to ask next is what happens when there is motion involved of either the source, the observer, or both.

Moving Source

The animation below shows the source moving toward a stationary observer. This source, for example, could be an speeding ambulance which is emitting a loud sound of a specific pitch. If you watch the animation closely, the frequency with which the source is emitting is different the frequency with which the observer is receiving them, the blue dot (the observer) is flashing more frequently than the red dot (the source). How is it possible that the observer could be measuring a frequency different from what the source is emitting?

Figure 8.4.2: Source Moving Toward Stationary Observer

Let us try to understand this phenomena from the perspective of the observer. It is apparent from the animation above, that the crests are more dense in front of the moving source and are less dense behind the source. Thus, the wavelength as measured by the observer in front of the moving source is shorter. When the source is stationary the distance between two consecutive crests is one *source wavelength*, λ_s, the wavelength as measured by the source. However, when the source is moving, every time it emits a crest it moves a certain distance before emitting the next crest. Thus, it gets closer to the crest in front of it when it emits the next

one, shortening the distance between two neighboring crests. Likewise, it gets further away from the crest behind it, lengthening the distance between crests behind the source.

We can figure out this new distance between the crests by calculating the distance that the source moves between consecutive emissions of crests. The time it takes to emit two neighboring crests is one *source period*, T_s, (period as measured by the source), the speed with which the source moves is *source speed*, v_s, so the distance it moves is $d = v_s T_s$. Thus, the new distance between two crests, which we call the *observed wavelength*, λ_o, in front of the moving source is the source wavelength minus the distance the source moved in one period:

$$\lambda_o = \lambda_s - v_s T_s \tag{8.4.1}$$

From this we can determine the frequency which the observer measures. We can use the relationships, $\lambda = v/f$ and $T = 1/f$, to rewrite equation 8.4.1 in terms of frequencies. We will leave the symbol, v, without any subscript to mean the speed of sound. Notice, the speed of sound is the same for both the source and the observer in this scenario since it depends on the medium and the observer is stationary (you will see shortly how the speed of sound changes according to a moving observer). Replacing wavelength with frequency in Equation 8.4.1 we get:

$$\frac{v}{f_o} = \frac{v}{f_s} - \frac{v_s}{f_s} \tag{8.4.2}$$

By rearranging the equation, we can express the observed frequency in terms of source frequency using the same method as above:

$$f_o = \frac{v}{v - v_s} f_s \tag{8.4.3}$$

You can see from the equation above that the observed frequency is larger than the source frequency when the source moves toward the observer, since the numerator, v is greater than the denominator, $v - v_s$. This is consistent with the animation of the blue observer dot flashing at a greater frequency compared to the red source dot.

If the observer is standing behind the moving source instead, the animation shows that the crests get farther apart from a receding source. In this case, the source moves away from an emitted crest by a distance of $v_s T_s$ before emitting the next crest. So the observed wavelength becomes:

$$\lambda_o = \lambda_s + v_s T_s \tag{8.4.4}$$

Resulting in the following expression for the observed frequency:

$$f_o = \frac{v}{v + v_s} f_s \tag{8.4.5}$$

In the equation above the observed frequency is lower than the source frequency, since the numerator, v is less than the denominator, $v + v_s$. Since the speed of sound is fixed, the shorter wavelength in front of an approaching source results in a higher frequency, while the longer wavelength behind a receding source results in a lower frequency. Next time an ambulance is zooming by you, try to pay attending to the pitch of the siren as it approaches you and then recedes away from you. This phenomena is known as the *Doppler effect*.

Combing the two equations, the Doppler effect for a moving source can be written as:

$$f_o = \frac{v}{v \pm v_s} f_s \tag{8.4.6}$$

where the sign is a plus for a source moving away from the observer and a minus for a source moving toward the observer.

Moving Observer

Let us consider what happens when, instead of the source, the observer is moving toward or away from a stationary source. In this case since the source is stationary the distance between the crests is the same in front and behind the source, as shown in the animation of Figure 8.4.1. Thus, in this scenario the wavelength is fixed. A moving observer would measure the same distance between crests as it would if it was stationary. If the blue dot representing the observer moves toward the source in Figure 8.4.1, it would flash more frequently since it would encounter crests more often, than it would if stationary. Thus, the frequency according to the observe increases. But how is this possible if the wavelength remains the same and the medium does not change?

Let us think of this in terms of an analogy. Imagine you are driving in a vehicle with a train on tracks parallel to the road approaching you. The observer is you in the moving vehicle, and the train represents the moving wave. The distance between the cars of the train appears the same to you, so the wavelength is the same. But the frequency with which you are seeing each new car as you are moving in the opposite direction of the train is greater than if you were stationary. The train also appears to be moving faster than it does when you are stationary. Of course, the train itself is moving at a constant speed, but according to the observer the train is moving faster. If on the contrary you were moving in the same direction as the train, the train would appear to move slower, or even stationary if your speed is identical to that of the train. This is known as *relative motion,* the speed of objects depends on the frame from which they are measured. A stationary observer will measure the train moving at a different speed from the observer in a moving vehicle.

Therefore, when the observer is moving toward the source emitting a wave, the speed of sound appears to increase. It does so according to this equation, $v+v_o$, where v the speed of sound according to the rest frame, and v_o the speed of the observer. Thus, the frequency as measured by the observer is the speed of sound according to the observer dividing by the wavelength emitted by the source:

$$f_o = \frac{v+v_o}{\lambda} = \frac{v+v_o}{v} f_s \tag{8.4.7}$$

where we used $\lambda = v/f_s$ for the wavelength. If the observer is moving away from the source, or analogously you are driving in the same direction as the moving train, then the speed measured by observer decreases, and the speed of sound according to the observer becomes $v-v_o$. Using the same reasoning as in Equation 8.4.7 the observed frequency for an observer moving away from the stationary source is:

$$f_o = \frac{v-v_o}{v} f_s \tag{8.4.8}$$

Combing the two equations, the Doppler effect for a moving observer and a stationary source is:

$$f_o = \frac{v \pm v_o}{v} f_s \tag{8.4.9}$$

The observed frequency is higher than the source one when the numerator is bigger than the denominator, which is the case when the sign is positive. This is consistent with the observation of the blue dot flashing more frequently if it was to move toward the source. On the contrary, we choose a negative sign when the observer is moving away from the source, resulting in a lower observed frequency.

Combined Doppler Effect

For completeness, let us consider the Doppler effect when both the source and the observer are moving. When both are moving, the observe measures a different speed of sound due to relative motion and a different wavelength due to source moving. Thus, the speed of sound measured by the observer is $v \pm v_o$ (depending on direction of motion) and the wavelength is λ_o. Thus, we get the following expression for the frequency:

$$f_o = \frac{v \pm v_o}{\lambda_o} \tag{8.4.10}$$

The wavelength that the observer measures is the same as calculated for a moving source. Using the result in Equation 8.4.6, we get a general expression for the combined Doppler effect:

$$f_o = \frac{v \pm v_o}{v \mp v_s} f_s \tag{8.4.11}$$

You can see from the equation above that when the observer is not moving, $v_o = 0$, we recover Equation 8.4.6 for the stationary observer and a moving source. When the source is stationary, $v_s = 0$, then we obtain Equation 8.4.9 for a stationary source and a moving observer. Naturally, if both are stationary, there is no Doppler effect resulting in $f_o = f_s$.

Since there are many parameters to keep track of when solving problems which involve the Doppler effect, here is a summary of all the terms in Equation 8.4.11:

- Observed frequency, f_o, frequency as measured by the observer.
- Source frequency, f_s, frequency as emitted by the source.

- Wave speed, v, is the speed of the wave as determined by the medium, most commonly applied to the speed of sound in air.
- Speed of the observer, v_o, use a plus sign in the numerator when the observer is moving toward the source and a minus sign when the observer is moving away from the source.
- Speed of the source, v_s, use a minus sign in the denominator when the source is moving toward the observer and a plus sign when the source is moving away from the observer.

Sound waves are not the only types of waves which exhibit the Doppler effect. In fact, the Doppler effect which was first described by an Austrian mathematician and physicist Christian Doppler in 1842 due to his observations of light coming from distant stars. The Doppler effect eventually led to the idea of an expanding universe, since it was observed that the light coming from distant stars always seems to be redshifted, meaning that the wavelengths of light emitted from the stars are longer than expected. This implies that the stars around us are always receding or moving away, consistent with a notion of an expanding universe. Even though, similar principles apply to light as to sound, since electromagnetic waves move at the speed of light, Doppler effect described by Equation 8.4.11 no longer applies. One needs to consider relativistic effects to calculate the Doppler effect.

There are also multiple applications of the Doppler effect in medicine, specifically in ultrasound technology. You will explore a model of ultrasound technology in Example 8.4.2 below. Police radars use the Doppler effect to calculate the speed of moving vehicles. The radar emits a wave, which then is reflected from the moving vehicle, and is detected by the radar. However, since radars emit electromagnetic waves, such as radio waves, relativistic Doppler effect equations need to be used.

Example 8.4.1

You are sitting in your car and you hear an ambulance which is moving toward you emitting a loud sound at a set frequency. Then you start moving in the same direction and at the same speed as the moving ambulance. Once you are moving you hear the frequency of the siren as 9/10 the frequency than you heard when you were stationary. Use the speed of sound as 344 m/s. Determine your speed.

Solution

When both you are the ambulance are moving in the same direction, the ambulance is moving toward you (so the sign in the denominator of the Doppler equation is negative) and the observer (you in the car) is moving away from the ambulance (so the sign in the numerator of the Doppler equation is also negative). Since both the car (the observer) and the ambulance (the source) are moving at the same speed, then $v_o = v_s$. Plugging this into the Doppler equation:

$$f_o = \frac{v - v_o}{v - v_s} f_s = \frac{v - v_o}{v - v_o} f_s = f_s$$

Thus, when the observer and the source are moving in the same direction and with the equal speed, there is no Doppler shift. To visualize this, think of driving on the highway right next to a car which is moving at the same speeds as you. That car appears to be stationary. Thus, when both the source and the observer are moving identically relative to each other, it is equivalent to both being stationary, resulting in no Doppler shift of frequency. Therefore, the source frequency (equivalent to the frequency the car observes when moving) is 9/10 the frequency the car observed when stationary, f_o:

$$f_s = \frac{9}{10} f_o$$

The ambulance is moving toward the car, so it observed a higher frequency when stationary. The Doppler equation for the scenario of a source moving toward a stationary observer is:

$$f_o = \frac{v}{v - v_s} f_s$$

Using the information of the ratio between the observed and source frequencies we get:

$$\frac{10}{9} = \frac{v}{v - v_s}$$

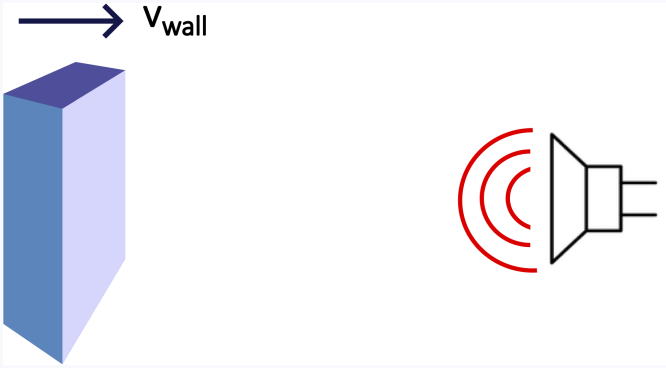
Solving for the speed of the source with some algebraic manipulations we obtain the following speed of the source:

$$v_s = \frac{v}{10} = \frac{344 m/s}{10} = 34.4 m/s$$

The above result is equivalent to the speed of the car once it starts moving as stated in the problem.

Example 8.4.2

Ultrasound technology uses the principle of Doppler effect to calculate the speed of blood in human veins. A sound wave generated by the ultrasound machine reflects off the moving blood cells and is then detected by the ultrasound machine. For simplicity, let us mimic this situation with a stationary sound source (a speaker replacing the ultrasound machine) and a wall moving toward the source (the wall replacing the moving blood), as seen below. Assume the source is emitting a frequency of 200 Hz, and the wall is moving with 5 m/s. Use the speed of sound in air as 344 m/s. Calculate the frequency of the sound that comes back to the source after the sound was reflected from the moving wall. Hint: you need to use the Doppler effect twice, once for the outgoing wave, and once for the reflected one.



Solution

When the wave is first emitted, the moving wall acts as an observer of this wave. It detects the frequency of this wave which can be calculated according to the equation of a moving observer toward a stationary source:

$$f_o = \frac{v + v_{\text{wall}}}{v} f_s$$

Once the wave is reflected, now the source detects that wave. The wall becomes the source of the wave, since it is now emitted the reflected wave. The new source frequency is then the observed frequency from the equation above. Using the Doppler effect equation for a moving source toward a stationary observer, the frequency of the reflected wave as detected by the original source is:

$$f_o = \frac{v}{v - v_s} f_s$$

The speed v_s is the speed of the wall, v_{wall}, and f_s is the frequency observed by the wall from the source as given in the first equation. Combining these results we get:

$$f_o = \frac{v}{v - v_{\text{wall}}} \cdot \frac{v + v_{\text{wall}}}{v} f_s = \frac{v + v_{\text{wall}}}{v - v_{\text{wall}}} f_s$$

Plugging in the given numerical values, we obtain the frequency of the reflected wave as detected by the source:

$$f_o = \left(\frac{344 + 5}{344 - 5}\right)(200Hz) = 205.9Hz$$

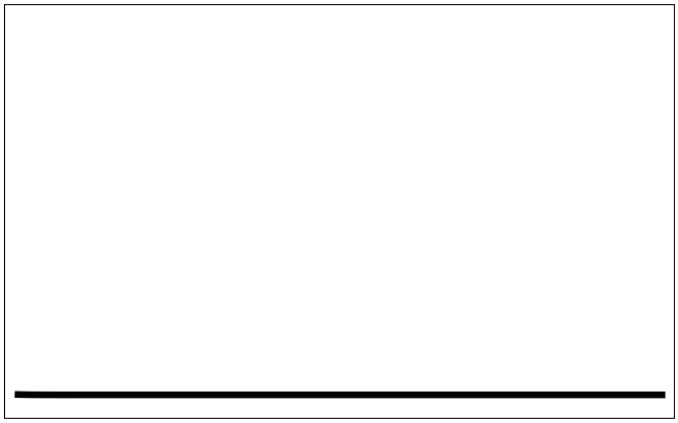
8.5: Superposition and Interference

Overview of Superposition

So far in our discussion of waves we analyzed the effect of one source creating a wave pulse, a repeating wave, or a harmonic wave. By knowing the motion and the energy input of the source, we can categorize the shape of the disturbance as it moves with speed determined by the medium. We also looked at the effect of the source and the observer moving. All of these discussions so far assumed one wave source present.

Now we want to explore what happens when two similar waves meet in space. An example of how this could occur is if you and a friend both held a rope. If you wiggle your end, the wave you make will propagate toward your friend. If your friend jiggles her end, her wave will propagate toward you. Another example is dropping two stones nearby in a pond. What happens when the waves collide? Eventually the ripples will overlap, but how can we calculate the resulting displacement from equilibrium? The addition of individual waves to obtain the total effect is called *superposition*. When waves meet at a given space and time their amplitudes simply add. In the animation below you can see two wave pulses coming together, their displacements adding when they overlap, and continuing to move in the same direction with their original amplitude once they are no longer overlapping.

Figure 8.5.1: Two Pulses Combining



This superposition can only occur with waves of the same type, such as water waves, sound waves, light waves, or waves on a string. Different types of waves can also meet in the same location in space, such as light and sound, but they will not interfere the same way since they are governed by different principles.

For two one-dimensional waves of the same type superposition can be written as:

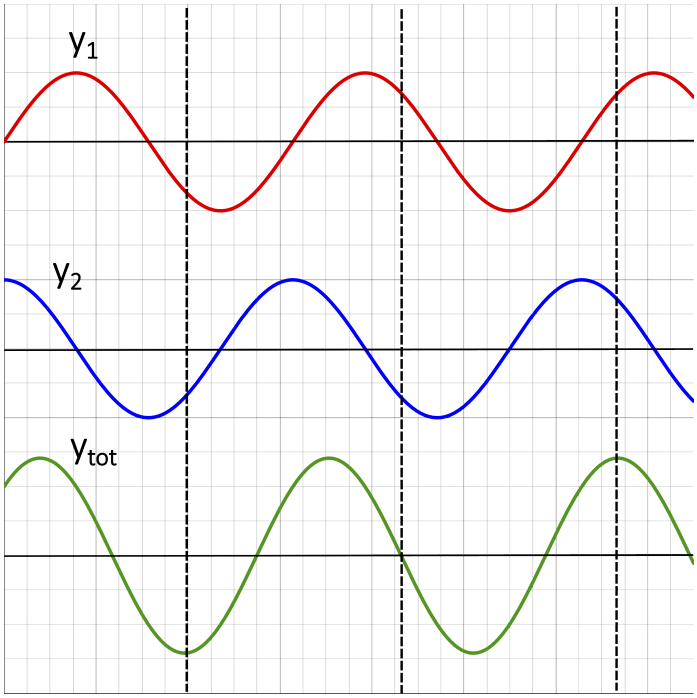
$$y_{\text{total}}(x,t) = y_1(x,t) + y_2(x,t) \tag{8.5.1}$$

The same principles apply to continuous harmonic waves, which specifically can be represented as:

$$y_{\text{total}} = A_1 \sin \Phi_1 + A_2 \sin \Phi_2 \tag{8.5.2}$$

Let us look at an example graphically of two waves with the same wavelength and amplitude meeting at the same location in space. Since these waves change with time, we can only look at a snapshot of this occurrence as shown in the figure below. Although, the two waves, y_1 and y_2 overlap in the same location, the wave are plotted separately for clarity.

Figure 8.5.2: Superposition

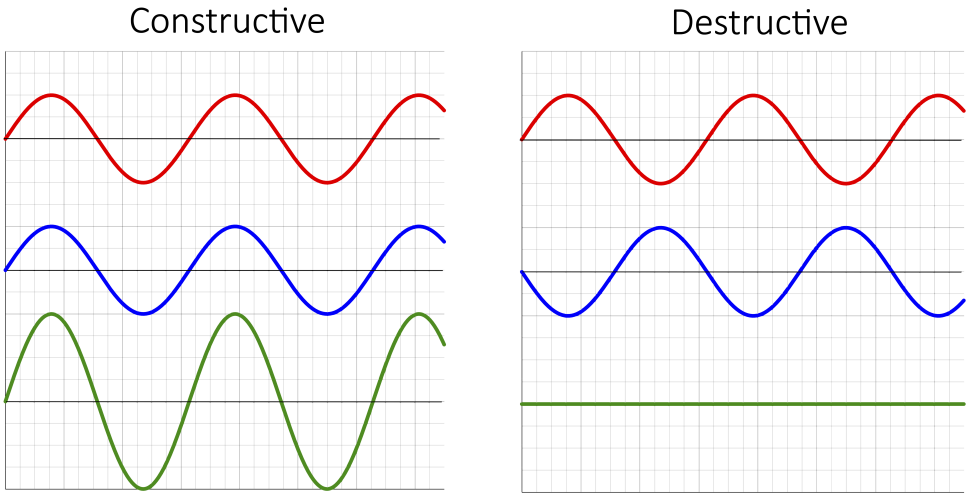


The two top plots are the two individual waves, y_1 and y_2, at some fixed time that meet in the same location in space. The bottom plot, y_{tot}, is the sum of the the two, which is obtained by summing the displacement of the two individual waves at each location. There are multiple locations marked with the vertical dashed lines which stress the addition of displacements. For example, at the middle dashed line the two displacements add up to zero, since y_1 is positive and y_2 is negative with the same magnitude of displacement. The amplitude of the total wave is clearly not the sum of the amplitudes of the individual ones, since the crests and the troughs do not overlap at any locations. The amplitude is bigger than that of the individual waves, since both waves have an amplitude of two grid lines and the amplitude of the combined wave is close to three grid lines. However, as we will see below this is not always the case. Sometimes the resulting amplitude is less then the either of the individual ones, or in the most extreme case it can be zero.

Special Cases of Interference

The main application of superposition for waves is *interference*. There are many ways in which any two sinusoidal functions can add, but there are two specific ones which are of special interest. In the left panel of the figure below two waves overlap perfectly. The displacements at each location are identical, so the amplitude of the superimposed wave is sum of the individual amplitude. This case is known as *constructive interference*. On the other hand, if one wave displaces up while other displaces down, the displacements add up to zero. This is the case of the two waves being half a cycle out of phase, and is known as *destructive interference*. As can be seen on the right panel of the figure below, the two waves add up to zero everywhere resulting in a complete cancellation of the waves. *Partial interference* is any kind of interference that isn't completely constructive or completely destructive, as in the example in Figure 8.5.2. The term partial interference is not very descriptive since we can have partial interference that is either almost constructive or almost destructive.

Figure 8.5.3: Constructive and Destructive Interference



As we can see from the examples in figures above, the type of interference that will occur depends on how the two wave line up relative to each other. When crests and troughs meet exactly, the waves are in phase and the resulting amplitude will be the sum of the individual ones. Thus, the total phases of the two waves are the same or different by a multiple of 2π, which keeps the sine function the same. In other words, the two phases in Equation 8.5.2 differ only by a factor of $2\pi n$. If the two initial waves have the same amplitude, then Equation 8.5.2 for the case of constructive interference becomes:

$$y_{\text{total}} = A\sin\Phi + A\sin(\Phi + 2\pi n) = 2A\sin\Phi \tag{8.5.3}$$

Thus, the new amplitude $2A$ is double the original one as see on the bottom combined wave in the left plot of Figure 8.5.3 above.

On the other hand, when a crest of one wave overlaps the trough of the other wave, the total phases of the two waves differ by an odd multiple of π resulting in the cancellation of the wave. This is depicted on the right panel of Figure 8.5.3. In this case Equation 8.5.2 becomes:

$$y_{\text{total}} = A\sin\Phi + A\sin\Big(\Phi + 2\pi(n+1/2)\Big) = A\sin\Phi - A\sin\Phi = 0 \tag{8.5.4}$$

where we have used an important property of the sine function, that shifting the phase of the sine function by half a cycle, negates the original function.

From these examples we can see that the difference in the total phase between the two waves determines the type of interference that we observe. For now we will focus on interference between two waves of the same frequency. Since waves interfere in the same medium, this implies that their wavelengths must be the same as well. For interference problems we will always measure position as increasing from the source. In other words, x is increasing with time, which is equivalent to a right-moving wave. Thus, we will use the minus sign in front of the spatial term in the expression for the total phase defined in a previous section:

$$\Phi = \frac{2\pi}{T}t - \frac{2\pi}{\lambda}x + \phi_o \tag{8.5.5}$$

We can see that there are three terms that can contribute to the difference in the total phase between two waves. Thus, a difference in any of the three terms could determine the difference in the total phase between two waves. Given that we are assuming sources of equal frequencies, a difference in the time-dependent term, $2\pi\Delta t/T$, could still arise if the interfering sources were turned on at different times. For example, assume that two speakers which are turned on at the same time and are in phase (have the same phase constant), such that they generate crests and troughs at the same time. If instead, you turned on one of the speakers half a period after the other on, then when one speaker is generating a crest the other one would be generating a trough, making them out of phase by half a cycle. However, this scenario is equivalent to two speakers which are turned on at the same time, but with a phase constant difference of half a cycle, $\Delta\phi_o = \pi$. Thus, you can always account for any difference in time that the waves were generated, as a difference in a phase constant. Therefore, for simplicity we will set the time difference term to zero, $\Delta t = 0$, by always synchronizing the time of the two sources. The total phase difference is then simplified to the following form:

$$\Delta\Phi = -\frac{2\pi}{\lambda}\Delta x + \Delta\phi_o \tag{8.5.6}$$

where the first term depends on the position of each source relative to the location where interference is measured (to be discussed in detail below), and the second term is the difference between the phase constants of the two sources.

The conditions for the the difference in the total phase for different types of interference are as follows:

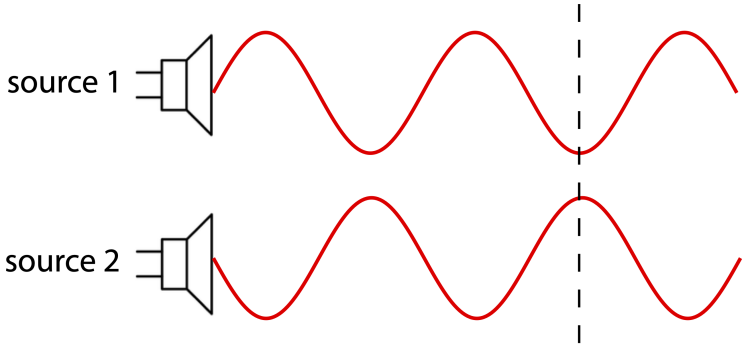
- **Constructive interference**: the total phase difference is any even multiple of π: $\Delta\Phi = 2\pi n$, where n is an integer ($n = 0, \pm 1, \pm 2, \ldots$).
- **Destructive Interference**: the total the phase difference is any odd multiple of π: $\Delta\Phi = 2\pi(n + 1/2)$, where n is an integer ($n = 0, \pm 1, \pm 2, \ldots$).
- **Partial interference**: total phase difference is anything other than the two conditions above.

Interference of Sound Waves

We will specifically analyze the effect of interference of sound waves. Since constructive interference results in the maximum amplitude obtained from summing two waves, in sound waves this will manifest itself as the loudest you can hear the sound from two sources. When the sound waves will interfere destructively resulting in a combined amplitude of zero, the sound will no longer be heard. As previously mentioned, for now we will only analyze sources of equal frequencies.

Equation 8.5.6 shows that there are two terms that could contribute to the total phase difference: a difference in location and a difference in phase constant. Let us look at effect of each one of these two terms. We start by looking at simple one-dimensional situations where the observer is standing to the right of two speakers. As in the example demonstrated in the figure below, we want to measure interference at the dashed vertical line to the right of the two sources. The two sources are vertically separated in the image for clarity, but physically the two waves are overlapping, and the two sources are at the same location (the speakers are placed side by side). In the snapshot shown below the waves coming out from both sources are at equilibrium. But a quarter of a cycle earlier, source 1 produced a crest while source 2 produced a trough. This means that the the two waves are half a cycle out of sync, implying that the difference in the phase constant is π. Graphically, we can see that crests overlap with troughs, resulting in destructive interference.

Figure 8.5.4: Phase Constant Difference



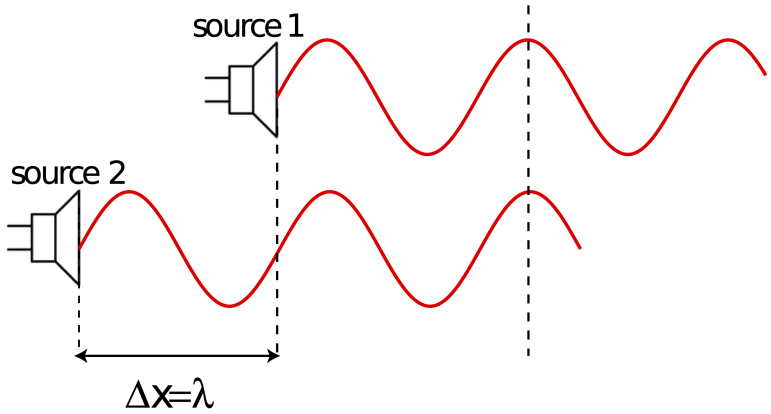
In this case since the two sources are at the same location, $\Delta x = 0$. The difference in their phase constants is π. This results in the following expression for the total phase difference:

$$\Delta\Phi = -\frac{2\pi}{\lambda} \cdot 0 + \pi = \pi \tag{8.5.7}$$

which confirms destructive interference which is observed in Figure 8.5.4. Note, this scenario could also be interpreted as two sources with equal phase constants being turned on half a period after one another. This would mean that there is a difference in time between the two sources, but as we already discussed, any difference in time can always be interpreted as a difference in a phase constant, such as we treated this problem here.

The figure below shows another example, where the phase difference is caused by difference in position of the two sources. Source 2 is behind source 1 by a distance of one wavelength. The two sources have the same phase constant, since both are producing identical waves in this snapshot. The only difference between the two sources is their relative positions.

Figure 8.5.5: Path Length Difference for Constructive Interference

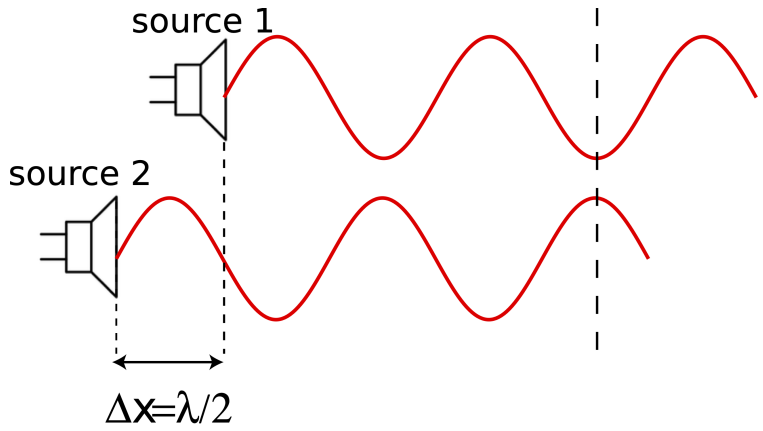


The position in this case matters since although the two speakers are in phase (equal phase constants), the wave from source 2 travels a longer distance to the observer compared to the wave from source 1. The difference between these two positions, Δx, is known as the *path length difference*. In this example, that path length difference is exactly one wavelength, $\Delta x = x_2 - x_1 = \lambda$, according to an observer who is standing to the right of both speakers (for example, at the location of the last vertical dashed line). This particular path length difference is identical to the situation if the two speakers were in the same locations. This results in constructive interference as seen above. Mathematically, the total phase difference for the situation in Figure 8.5.5 is written as:

$$\Delta\Phi = -\frac{2\pi}{\lambda}\lambda + 0 = -2\pi \tag{8.5.8}$$

giving the condition for constructive interference. The figure below shows a similar situation, except now source 2 is half a wavelength behind source 1, such that $\Delta x = x_2 - x_1 = \lambda/2$. As can be seen in the figure, in this case crests and lining up with troughs, which will resulting cancellation of amplitudes.

Figure 8.5.6: Path Length Difference for Destructive Interference

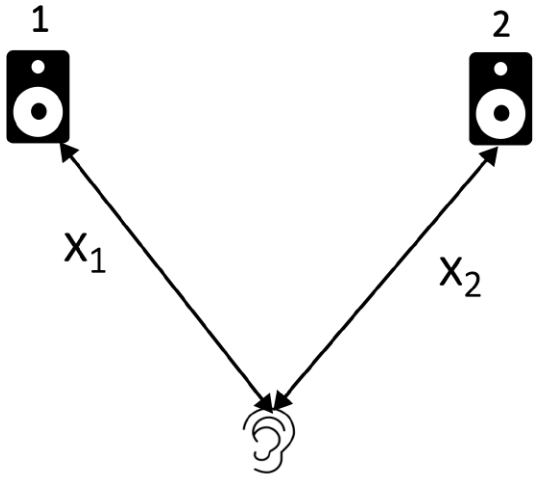


Mathematically, plugging in for the the path length difference we obtain the result for destructive interference:

$$\Delta\Phi = -\frac{2\pi}{\lambda}\frac{\lambda}{2} + 0 = -\pi \tag{8.5.9}$$

Notice that the difference is positions stays the same everywhere on the right of the two speakers. Although, we choose to draw the waves only emanating to the right of the speakers, in reality they are being generated in three-dimensions around the speakers. To the left of the two speakers, for example, the value of, $x_2 - x_1$, would be negated since now source 2 is now closer. But if interference is being measured somewhere where the speakers and the observer no longer make a straight line, the same concepts hold for the path length difference. The illustration below shows a general description of path length difference.

Figure 8.5.7: Path Length Difference



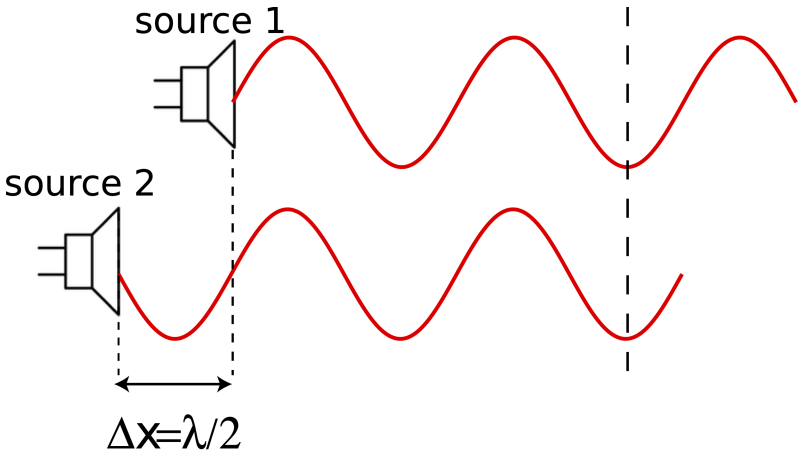
The observer is standing a distance x_1 from speaker 1 and a distance of x_2 from speaker 2. The path length difference is then defined as:

$$\Delta x = x_2 - x_1 \tag{8.5.10}$$

If the interference is being measured locations which are the same distance to each speaker, then the path length difference is zero. Otherwise, it is important to calculate what fraction of the wavelength is the difference in the distance to each source. Note, once you define one particular source as "1" and the other as "2", make sure to be consistent in calculating the total phase difference, such that you are subtracting the phase constant of speaker 1 from the phase constant speaker 2, as well.

The last example shown below depicts a combination of the effect of a difference in both the phase constant and the path length. The two sources shown have a phase constant difference of π, and source 2 is half a wavelength behind source 1.

Figure 8.5.8: Path Length Difference and Phase Constant Difference



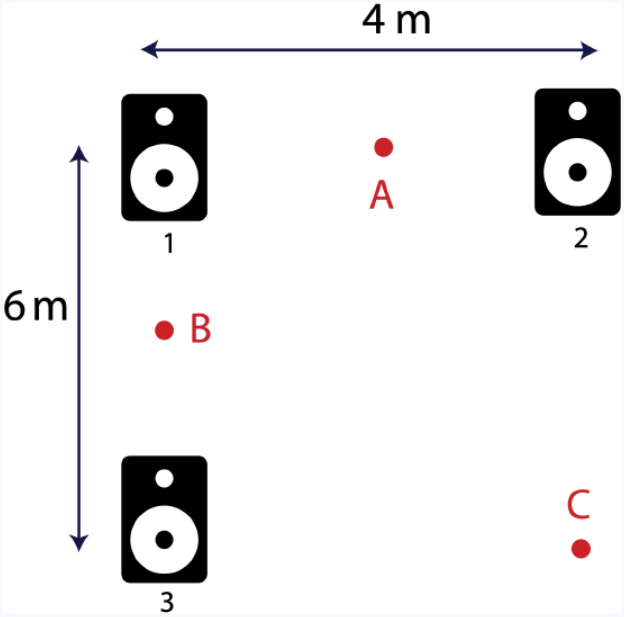
Plugging this scenario into the equation for the total phase difference, we obtain:

$$\Delta\Phi = -\frac{2\pi}{\lambda}\frac{\lambda}{2} + \pi = 0 \tag{8.5.11}$$

which gives the condition for constructive interference, as can be seen in the figure since crests and troughs are lining up from each wave. This example stresses that it is important to consider the combination of both effects. We saw in the example presented in Figure 8.5.4 that a phase difference of half a cycle results in destructive interference. Then we saw from Figure 8.5.6 that a path length difference of half of a wavelength also results in destructive interference. In this example, both of these effects, which individual give rise to destructive interference, when are presents simultaneously result in constructive interference.

Example 8.5.1

In the diagram below there are three speakers. When turned on all three are emitting a tone with a wavelength of 2 m. Alice who is standing at location A hears no sound when only speakers 1 and 2 are turned on. Brendan who is standing at location B hears a loud sound of maximum intensity when only speakers 2 and 3 are turned on.



Calculate $\Delta\Phi$ and determine the type of interference that Crystal standing at location C will hear when:

a) Only speakers 1 and 2 are turned on.

b) Only speakers 1 and 3 are turned on.

c) Only speakers 2 and 3 are turned on.

Solution

First we need to use the provided information to figure out the phase constant differences between the various speakers. When speakers 1 and 2 are turned on, the interference is destructive at location A: $\Delta\Phi_{12} = 2\pi(n+1/2)$. *The path length difference is zero,* $\Delta x_{12} = 0$, *since Alice is equidistant to speakers 1 and 2. This results in the following expression for the fixed phase constant difference between the two speakers,* $\Delta\phi_{12}$*:*

$$\Delta\Phi_{12} = 0 + \Delta\phi_{12} = 2\pi\left(n + \frac{1}{2}\right)$$

Choosing the $n = 0$ *solution for convenience, we obtain the following expression for the phase constant difference between speakers 1 and 2:*

$$\Delta\phi_{12} = \pi$$

When speakers 2 and 3 are turned on the interference is constructive at location B, $\Delta\Phi_{23} = 2n\pi$, *since Brendan hears maximum loudness. The distance from B to speaker 3 is 3 m. To find the distance to speaker 2, we need to use Pythagorean theorem, since point B, speaker 1, and speaker 2 make a right triangle. The distance from B to speaker 2 is the hypotenuse of this triangle:*

$$x_2 = \sqrt{3^2 + 4^2} = 5\ m$$

Putting these results together to obtain the difference in the total phase at location B when speakers 2 and 3 are turned on:

$$\Delta\Phi_{23} = -\frac{2\pi}{2m}(5-3) + \Delta\phi_{23} = 2\pi n$$

Simplifying and choosing the $n = -1$ *solution we find:*

$$\Delta\phi_{23} = 0$$

Since the speakers 3 and 2 are in phase, the phase difference between 1 and 2 will be the same as the phase difference between 1 and 3:

$$\Delta\phi_{13} = \pi$$

Now, we ready to answer the three questions about inference that Crystal hears when all the possible pairs of speakers are turned on. The distance to speaker 1 from location C is obtained using Pythagorean theorem again:

$$x_1 = \sqrt{4^2 + 6^2} = 2\sqrt{13}m$$

The distance to speaker 2 is 6 m, $x_2 = 6m$, and the distance to speaker 3 is 4 m $x_2 = 4m$.

a) When speakers 1 and 2 are turned on the path length difference is, $x_1 - x_2 = (2\sqrt{13} - 6)$ m. The difference in the total phase is:

$$\Delta\Phi_{12} = -\frac{2\pi}{\lambda}\Delta x_{12} + \Delta\phi_{12} = -\frac{2\pi}{2m}(2\sqrt{13} - 6)m + \pi = -0.211\pi$$

The interference is partial and is closer to constructive since the phase difference is closer to zero than to π.

b) When speakers 1 and 3 are turned on the path length difference is, $x_1 - x_3 = (2\sqrt{13} - 4)$ m. The difference in the total phase is:

$$\Delta\Phi_{13} = -\frac{2\pi}{\lambda}\Delta x_{13} + \Delta\phi_{13} = -\frac{2\pi}{2m}(2\sqrt{13} - 4)m + \pi = -2.211\pi$$

The total phase difference above is identical to part a) since the two results differ by 2π, so the interference is the same due to these first two scenarios.The only difference between the two cases is the distance to speaker 2 is 2m longer than the distance to speaker 3. But this does not make a difference when it comes to interference since 2m is exactly one wavelength, so the path length difference is difference by one wavelength in part a). One wavelength is a shift by one cycle, resulting in identical interference.

c) When speakers 2 and 3 are turned on the path length difference is, $x_2 - x_3 = 6 - 4 = 2m$. The difference in the total phase is:

$$\Delta\Phi_{23} = -\frac{2\pi}{\lambda}\Delta x_{23} + \Delta\phi_{23} = -\frac{2\pi}{2m}(2m) + 0 = -2\pi$$

Thus, the inference in this case is constructive.

Example 8.5.2

You happen to be lost in a middle of a desert, in between two radio towers which are 15 km apart. Both towers are transmitting radio waves of frequency $3.75 \times 10^4\ Hz$. The radio signals are interfering with your GPS and not allowing you to find your way out. When you are standing along the straight line between the two towers 7 meters away from tower A, you find constructive interference. Instead, you want to find locations of destructive interference, so you can use your navigation system. Find all such positions on the line connecting towers A and B. Express your answer in terms of distance from tower A. Note, radio waves are electromagnetic waves which travel at the speed of light, $c = 3 \times 10^8\ m/s$. Electromagnetic waves also interfere when they meet in the same location in space, using the same principles as sound waves.

Solution

The goal here is to find all possible path length differences between towers A and B where interference is destructive. The phase constant difference is not given directly, but we are told that the interference is constructive when you are 7 km away from tower A. At this location you are 8 km from tower B, since they are 15 km apart. Therefore, the path length difference is:

$$x_B - x_A = 8 - 7 = 1km$$

We also need to determine the wavelength in order to calculate the total phase differences. The wavelength for these radio waves is:

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8 m/s}{3.75 \times 10^4 Hz} = 8000\ m = 8\ km$$

Since the interference is constructive at this location, the total phase difference is $2\pi n$:

$$\Delta\Phi_{BA} = -\frac{2\pi}{\lambda}(x_B - x_A) + \phi_B - \phi_A = -\frac{2\pi}{8km}(1km) + \phi_B - \phi_A = 2\pi n$$

Solving for the fixed phase constant difference:

$$\phi_B - \phi_A = 2\pi n + \frac{\pi}{4} = \frac{\pi}{4}$$

Note, we chose $n = 0$ solution for convenience.

Now that we know the phase constant difference, let us find all the path length differences between the two towers which will result in destructive interference:

$$\Delta\Phi_{BA} = -\frac{2\pi}{8}(x_B - x_A) + \frac{\pi}{4} = 2\pi\left(n + \frac{1}{2}\right)$$

Solving for path length difference:

$$x_B - x_A = -8n - 3$$

The equation above has an infinite number of solutions, but we are constrained since your locations needs to be on the line connecting the two towers. The closest possible position to tower A is if you were standing at the location of A, resulting in $x_A = 0$ and $x_B = 15$. The furthest position from tower A is at tower B, $x_A = 15$ and $x_B = 0$. Therefore, the path length difference can only have the following values between the two towers:

$$-15 \le (x_B - x_A) \le 15$$

Thus, we need to find all values of n which satisfy the above constraint. Starting with small values of n and increasing them until we no longer satisfy the above equation, we get the following four solutions:

$$\begin{aligned} n = 0: &\quad x_B - x_A = -3\ km \\ n = 1: &\quad x_B - x_A = -11\ km \\ n = -1: &\quad x_B - x_A = 5\ km \\ n = -2: &\quad x_B - x_A = 13\ km \end{aligned}$$

A positive value of $n \ge 2$ or a negative value $n \le -3$ would result in a path length difference outside the allowed range, placing you in either the right or the left of both towers. The path length difference does not give you the exact distance from tower A, x_A. But we can use the fact that the sum f the distances to towers A and B has to be 15 km:

$$x_A + x_B = 15\ km$$

Now we have two equations with two unknowns, so we can solved for x_A for each of the four solutions above:

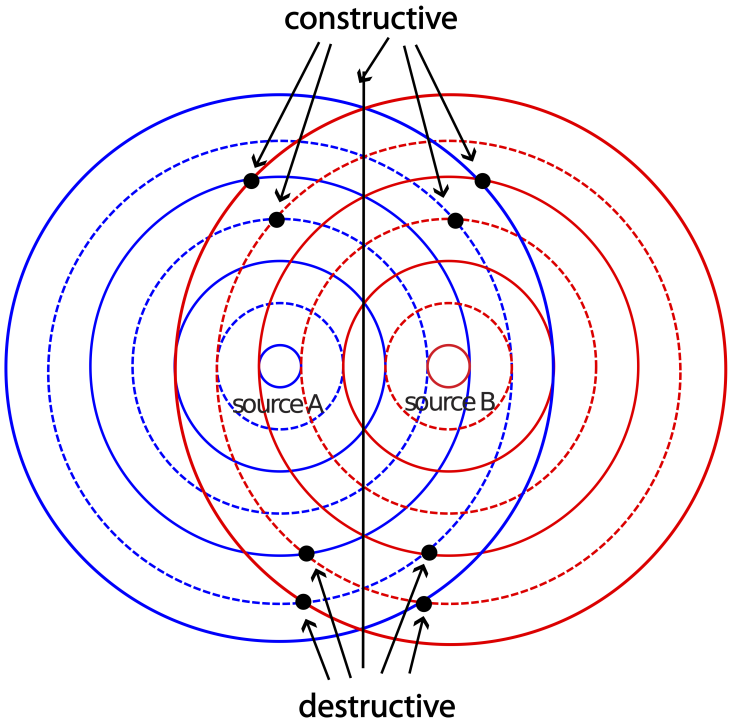
$$x_A = 1\ km, 5\ km, 9\ km, \text{and } 13\ km$$

To check that these values are correct, plug them back into the phase difference equation and check that you get destructive interference.

Wavefront Representation of Interference

We can also use the wavefront representation as another tool of depicting wave interference. Shown in the figure below are two sources generating waves with equal wavelengths, and thus frequencies. The crests (solid lines) and trough (dashed lines) emitted by source A are blue, while those emitted by source B are red. The two sources are near each other such that their wavefronts will meet and overlap in the same location in space.

Figure 8.5.9: Interference between Two Wavefronts

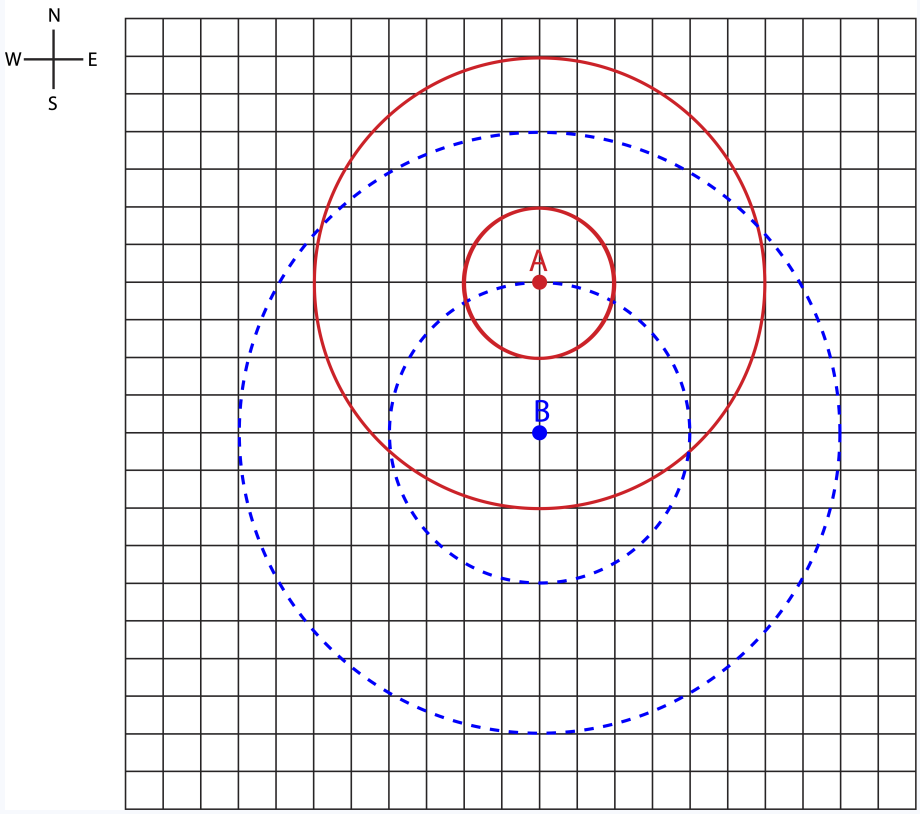


From what we have learned about interference, the two waves will interference constructively when two crests or two troughs overlap. Some of these locations are marked in the figure above. The central line separating the two sources shows a continuous location of constructive interference. How is it possible to conclude that since there are only discrete locations of constructive interference that can be interpreted from the figure due to overlapping crests and troughs? The two sources shown are in phase, since crests and troughs are being being generated at the same time. (On the contrary, if you were to replace all the crests for one of the sources by troughs and all the troughs by crests, that would depict two sources which are half a cycle of of phase.) In addition, the path length difference is zero along the central line since the distances to each source are equal. This implies that the total phase difference along the central line will always be zero, resulting in constructive interference.

Locations of destructive Interference are marked at some of the places where a crest from one source meets a trough from the other source, resulting in cancellation of amplitudes. The picture above is a snapshot at a particular instance of time. At a later time, there will be more wavefronts present and more visible locations of constructive and destructive interference. The wavefront representation does not give us a full picture of interference, since we only have information about crests and troughs at a given instance of time, but it provide a good guide. Unlike the central line which is a highly symmetric situation, it is more difficult to predict all the locations of interference without doing a precise calculation of path length differences.

Example 8.5.1

Two sources, A and B, are emitting sounds waves. The picture below shows two crests which originated from source A (solid red circles) and two troughs from source B (dashed blue circles). Assume everything is depicted to scale.



a) Determine the type of interference (constructive, destructive, or partial) directly North of both sources, directly South of the sources, and along the horizontal line midway between the two sources. Explain your reasoning.

b) If you decreased the distance between the same two sources by half, which of the answers from part a) would change and how?

Solution

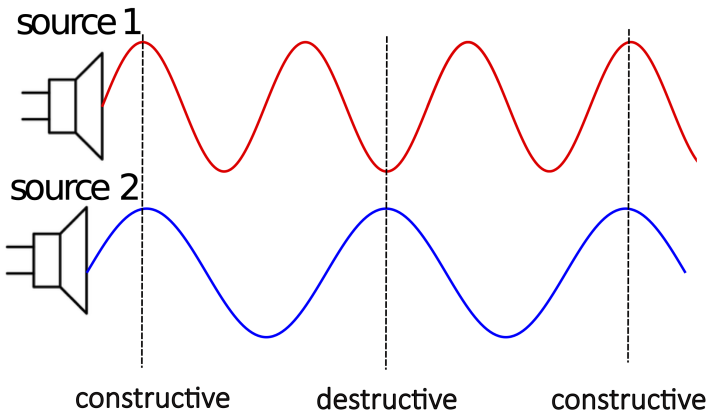
a) The wavelength is 4 units (distance between 2 crests or two troughs). The distance between the two sources is also 4 units. Thus, the two sources are one wavelength apart. It it not shown in the figure, but there is a trough for source A halfway between the two crests, or a distance of 4 units from source. That's exactly the distance of a trough emitted from source B. Therefore, when source A is emitting a trough, source B is emitting a trough as well, so the two sources are in phase (the difference between their phase constants is zero). On the North and the South of the two sources, the path length difference is just the separation between the sources. Therefore, North or South of the two sources the interference is constructive (path length difference is one wavelength, and the sources are in phase). Along the line midway between the two sources the interference is also constructive (path length difference is zero and the sources are in phase).

b) The new distance between the sources is half a wavelength, so the path length difference for North and South change to $\lambda/2$, making the interference destructive there. The midpoint still has zero path length difference, so the interference does not change.

8.6: Beats

So far we have focused on interference between two waves with the same frequency, and thus, wavelength. Interesting phenomena emerges when two waves with different frequencies interfere. To analyze this, let us start with a simplified one-dimensional situation of observing interference to the right of two speakers. In this case, the wavelengths of the two waves are different, since their frequencies are different as depicted in the figure below.

Figure 8.6.1: Wave Interference with Different Frequencies

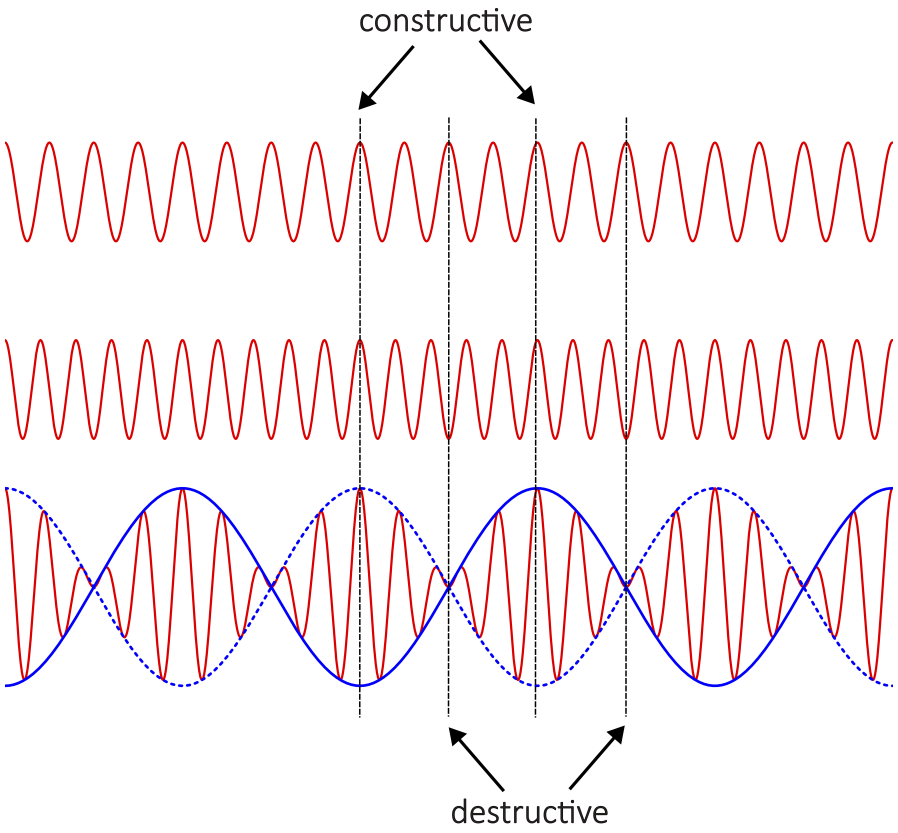


The figure above shows that interference is no longer fixed to the right of the speakers. In the previous section we determined that in this one-dimensional situation the type of interference is fixed for a given position, for which one can determine the difference in phase constants and path length between the two speakers. In other words, for a specific speaker set-up an observer on the right of the speaker always heard either maximum loudness, silence, or something in between. In this case, however, there are locations of destructive, constructive, and partial interference to the right of the speakers. The diagram above is a snapshot of the two waves at some particular time. If you are standing at the most right constructive interference location marked in the figure, you hear a sound of maximum intensity. However, at the later time, since the waves are traveling to the right, the location in the middle where interference is destructive will reach the location of the constructive interference on the right, so the observer will hear no sound. And some time even later, the most left location of constructive interference will reach the observer's location, so they will hear a loud sound again.

Unlike for the case of equal frequencies, we conclude that an observer at a fixed location now hears the intensity of sound changing periodically. In other words, the combined amplitude leading to a specific type of interference is now time-dependent. We call this phenomena *beats*, since the sound is now literally beating from loud to silent and back to loud, like a drum being hit in a periodic manner. This phenomena is especially relevant for two frequencies which are very closer together, since only then the beats can be perceived by humans. When the difference between the frequencies is large, the beats become too frequent for our ears to distinguish them.

Since we are combining two wave equations with different frequencies resulting in a time-dependent amplitude, the combined wave can no longer be described by a simple sinusoidal wave equation. However, it is still periodic since the amplitude change is periodic. The figure below shows an example of a combined wave from two sinusoidal waves with two different frequencies. The two top plots represent the two individual waves, while the bottom one is the combined one.

Figure 8.6.2: Beats



The blue *envelope wave* on the combined plot outlines the change in amplitude. One can see when the envelope function is at peak, the two individual wave overlap at the same displacement (crests in this figure) resulting in constructive interference at that location. When the envelope function is at equilibrium, the displacements of the individual waves are cancelling (one is at a crest while the other one is at a trough) resulting in destructive interference. This is a snapshot of the combined wave at a fixed time. If you are standing at a fixed position, this wave will move in time, so you will be hearing a periodic intensity of constructive and destructive interference. The period of this time-dependent interference, or beats, is half the period of the envelope function drawn above. The intensity of the sound is the square of the amplitude, so constructive interference (maximum intensity) occur every time the envelope wave is at a crest or a trough, or every half a cycle of the envelope function. Likewise, the envelope function has two locations of zero amplitude for every cycle. The frequency with which beats are heard is known as the *beat frequency*. For example, a beat frequency of 1 Hz means that the observer is hearing a beat (a loud sound) one time every second.

We can also determine the beat frequency mathematically in terms of the original frequencies of the two waves. The total phase difference between two waves with different frequencies and wavelengths can be written as:

$$\Delta\Phi = 2\pi t\Delta\left(\frac{1}{T}\right) + 2\pi\Delta\left(\frac{x}{\lambda}\right) + \Delta\phi_o \tag{8.6.1}$$

To determine the beat frequency, we want to find the time it takes for the interference pattern to repeat itself. Let us assume that at some specific time an observer standing at a fixed location hears constructive interference (the loudest sound from the combined waves). Let us define this time as $t = 0$ sec. Therefore, Equation 8.6.1 becomes:

$$\Delta\Phi = 2\pi\Delta\left(\frac{x}{\lambda}\right) + \Delta\phi_o = 0 \tag{8.6.2}$$

At time later time which we define as the beat period, T_b, the observer will again hear constructive interference. This implies that the phase difference increased from 0 to 2π. The path length difference and the phase constant difference is not time-dependent as long as the observer stays put, so the sum of the second two terms in Equation 8.6.1 remain zero at $t = T_b$. Therefore, the total phase difference at $t = T_b$ is:

$$\Delta\Phi = 2\pi T_b \Delta\left(\frac{1}{T}\right) = 2\pi T_b\left(\frac{1}{T_1} - \frac{1}{T_2}\right) = 2\pi \tag{8.6.3}$$

Re-writing the above equation in terms of frequency and solving for the beat frequency, we get:

$$2\pi\frac{f_1 - f_2}{f_b} = 2\pi \Longrightarrow f_b = |f_1 - f_2| \tag{8.6.4}$$

The absolute value in the equation above implies that the beat frequency which is the difference between two frequencies is a positive value, since frequency cannot be negative. Thus, the absolute value assure that the the beat frequency is positive, if you happen to subtract the larger frequency from the smaller one. You can see from this results that the difference in the two frequencies needs to be small for beats to have a meaningful interpretation to the human ear. It is very difficult to distinguish even 10 individual beats in a short time span of one second.

Another important property of beats is the *carrier frequency*. This frequency is approximately the frequency of the "small" oscillations within the envelop function. In other words, in combined plot of Figure 8.6.2, the carrier period would be the time between two consecutive peaks (approximately). The carrier frequency, f_c, is the pitch you hear coming from the sound which is beating, and it turns out to be simply the average of the two original frequencies:

$$f_c = \frac{f_1 + f_2}{2} \tag{8.6.5}$$

Digression: Beat Wave Equation

We can mathematically determine that beat and carrier frequencies by adding two wave functions of waves with different frequencies. Recall, for sound waves we represent the displacement in terms of pressure, with equilibrium displacement at atmospheric pressure, which we set here to zero for simplicity. The two wave function can be written as:

$$P_1(x,t) = P_o \sin\left(\frac{2\pi}{T_1}t - \frac{2\pi}{\lambda_1}x + \phi_1\right), \quad P_2(x,t) = P_o \sin\left(\frac{2\pi}{T_2}t - \frac{2\pi}{\lambda_2}x + \phi_2\right)$$

Let us assume that at some location and instance of time which we define as x=0 m and t=0 sec the interference is constructive, such that there is no relative phase constant difference between the two waves, $\phi_1 = \phi_2 = 0$. We want to see how the wave changes as a function of time at any position, so we choose x=0 m for convenience. Also, we will express the wave functions in terms of frequency instead of period. Thus, the sum of the two wave functions simplifies to:

$$P_{tot}(t) = P_o \sin(2\pi f_1 t) + P_o \sin(2\pi f_2 t)$$

Here we need to use a trigonometric identity:

$$\sin\alpha + \sin\beta = 2\cos\frac{\alpha - \beta}{2}\sin\frac{\alpha + \beta}{2}$$

Applying this identify, we then get:

$$P_{tot}(t) = 2P_o \cos\left(2\pi\frac{f_1 - f_2}{2}t\right)\sin\left(2\pi\frac{f_1 + f_2}{2}t\right)$$

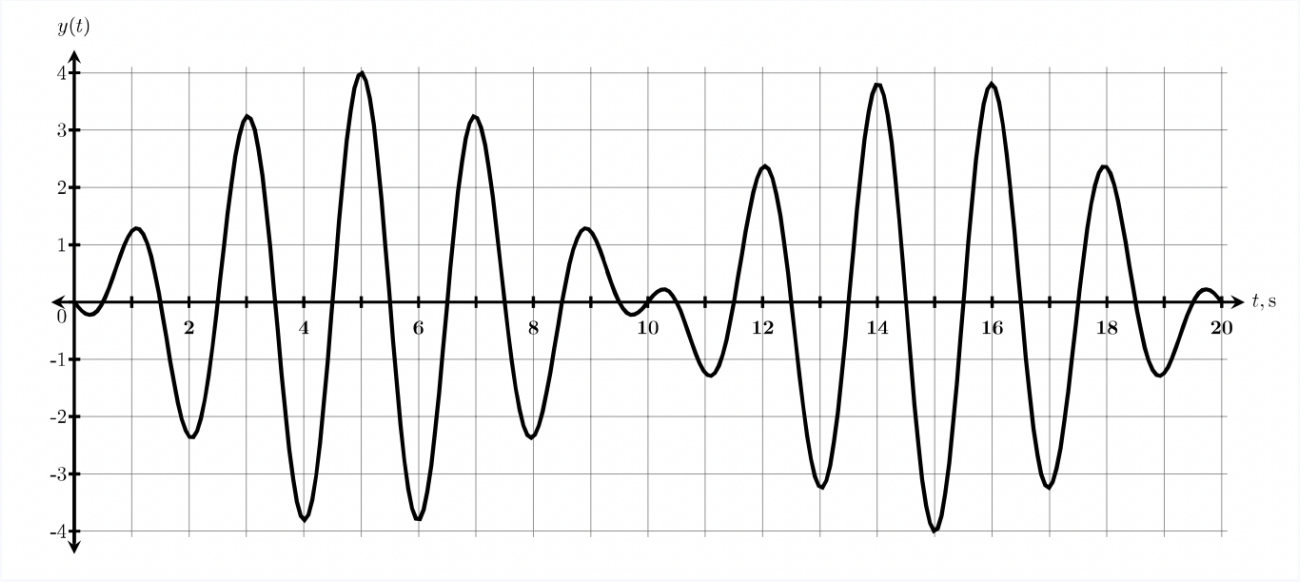
The above function is not the simple wave function with one sine term that we are familiar with, but represents the combination of the two waves. We can think of the second sine function as representing the carrier wave with the carrier frequency of $f_c = (f_1 + f_2)/2$. The quantity $2P_o$ multiplied by the cosine function represents the time-dependent amplitude of the combined wave. From the equation, we can be that frequency of this amplitude is $(f_1 - f_2)/2$. The beat frequency is twice this value, since for each cycle of the envelope function there are two beats, one crest and one trough both representing maximum intensity. This results in the same expression for beat frequency we obtain in Equation 8.6.4, $f_b = f_1 - f_2$. The maximum amplitude is $2P_o$ which exactly corresponds to constructive interference when the amplitudes of the two original waves are added.

Musicians use the principles of beats to tune their instruments. Frequencies on strings which are confined between two ends, like a guitar string, are determined mainly by the length and the speed of the wave on the string, as we will see in a later section on standing waves. The source no longer has control of the frequencies that it generates on the string, precisely because the ends of the string are confined. This is very different from what we have learned so far, so stay tuned! When tuning a string instrument, the

musician changes the tension of the string, which in turn changes the speed and results in a different frequency. They find the desired frequency by using tuning forks (at least this was the main method before tuning apps). A tuning fork has a predetermined frequency based on its physical properties at which it generates a sound when someone strikes it. Thus, with multiple turning forks of different frequencies a musician can tune all the strings to their desired frequencies. They do this by hitting the tuning fork and plucking or striking the string of the instrument at the same time. If the musician hears a beat, the string is out of tune. Then, they will adjust the tension until beats are no longer heard, implying that the string is vibrating at the exact same frequency as the tuning fork. Since the string which is out-of-tune is typically very close to the desired frequency, this results in a small beat frequency which can be easily heard by a human ear.

Example 8.6.1

The graph below shows a combination of two waves interacting at a particular point in space. Determine the frequencies of each wave.



Solution

The plot above is the sum of the two waves with different frequencies. The carrier frequency can be found from the period of the small oscillations inside the beat envelope. From the plot we see that the distance in time between any two consecutive peaks is 2 seconds, resulting in a carrier frequency:

$$f_c = \frac{1}{T_c} = \frac{1}{2s} = 0.5Hz = \frac{f_1 + f_2}{2}$$

The beat period can also be obtained from by measuring the time that passed between locations of biggest amplitude. This location does not have to be a crest, but can also be a trough, since the intensity of a wave is proportional to the square of amplitude. In this plot, we see the largest amplitude at t=5 sec. The next time that we observe the same amplitude is at t=15 sec. Therefore, the beat period is 10 sec, and the beat frequency is expressed as:

$$f_b = \frac{1}{T_b} = \frac{1}{10s} = 0.1Hz = f_1 - f_2$$

where we have assume that the bigger frequency is f_1. Now we can solve for the individual values of frequency, f_1 and f_2, since we have two equations and two unknowns. Plugging in beat frequency into the carrier frequency equation we get:

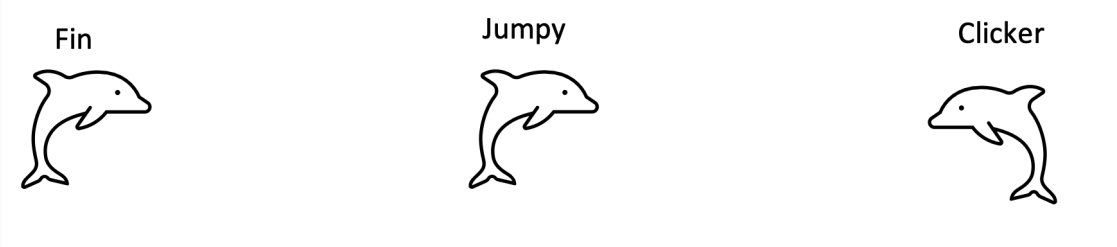
$$f_c = 0.5Hz = \frac{f_1 + f_2}{2} = \frac{f_1 + f_1 - 0.1}{2} = f_1 - 0.05 \Longrightarrow f_1 = 0.5 + 0.05 = 0.55Hz$$

Solving for the other frequency using the beat frequency equation:

$$f_2 = f_1 - 0.1 = 0.55 - 0.1 = 0.45Hz$$

Example 8.6.2

Three dolphins, Fin, Jumpy, and Clicker are communicating in water by emitting sounds of various frequencies. Fin and Clicker are both stationary. They are emitting a sound of the same frequency of 5000 Hz, while Jumpy is not emitting a sound. As Jumpy moves toward Clicker and away from Fin, he hears a beat frequency of 40 Hz. The speed of sound in water is 1500 m/s. Calculate Jumpy's speed.



Solution

If Jumpy was stationary, he would hear the same frequency coming from both Clicker and Fin, and therefore, would not hear any beats. But since Jumpy is moving toward and away from a sound source, his perceived frequencies of both sources are Doppler shifted, resulting in beats. The observed frequency by Jumpy of Clicker's sound, f_{oC}, is due to the observer moving, Jumpy moving at speed v_J, toward a stationary source with frequency emitted by source Clicker, f_C:

$$f_{oC} = \frac{v + v_J}{v} f_C$$

The observed frequency by Jumpy of Fin's sound, f_{oF}, is due to the observer moving away from the stationary source with frequency emitted by source Fin, f_F:

$$f_{oF} = \frac{v - v_J}{v} f_F$$

Setting the source frequencies equal to each other, $f_C = f_F \equiv f_s$, and subtracting the two equations to determine the beat frequency we get:

$$f_b = f_{oC} - f_{oF} = \frac{v + v_J - v + v_J}{v} f_s = \frac{2 v_J f_s}{v}$$

Solving for Jumpy's speed:

$$v_J = \frac{f_b v}{2 f_s} = \frac{40 Hz \times 1500 m/s}{2 \times 5000 Hz} = 6 \frac{m}{s}$$

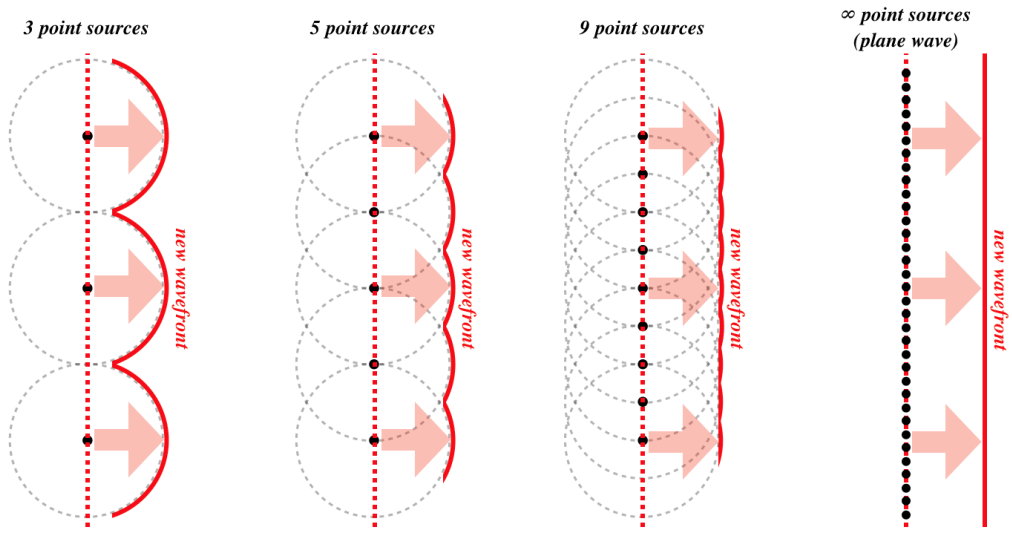
8.7: Double-Slit Interference

Huygens' Principle

Before diving into the the double-slit phenomena, we introduce a new tool that can be used to explain a lot of wave phenomena, known as the *Huygens' Principle.* It is named after a Dutch scientist Christiaan Huygens. In 1678 Huygens proposed that each point along a moving wavefront can be represented by a point source of a spherical waves. The wavefronts that are observed are due to the interference of all of these spherical waves.

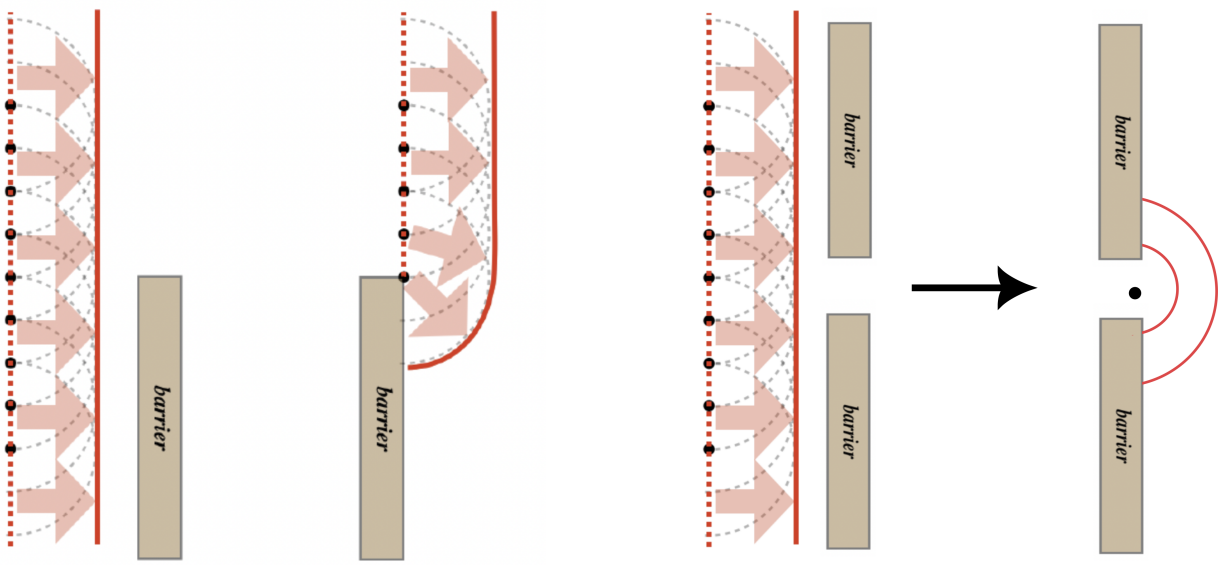
The figure below demonstrates Huygen's Principle by starting with a simplified picture of just three point sources on the very left of the figure. These three waves will interfere giving rise to the "new wavefront" which is wavelike. As you add more and more *wavelets* close together along a line, the overlap between the wave becomes more and more regular, until in becomes continuous giving rise to a straight wavefront when we have an "infinite" number of sources on the very right of the figure. Therefore, a plane wave coming from a distant object can be described as a combination of infinite number of point sources each generating an independent wavelet.

Figure 8.7.1: Huygens' Principle



Huygen's Principle is especially useful when explaining what happens to wavefronts when they encounter barriers. In the figure below on the left, a wavefront represented by a superposition of infinite point sources, encounters a barrier. The part of the wavefront that hits the barrier gets either reflected or absorbed by its surface, while the rest of the wavefront propagates forward. However, the wavelet generated by the point source right at the boundary of the barrier will no longer interfere with neighboring point sources below it. Thus, the spherical waves generated from the point sources near the barriers will be retained as shown in the figure. This will result in the wave bending around the barrier as shown in the figure. Recall, the direction of wave propagation is always perpendicular to the wavefront, as described in the ray representation in an earlier section.

Figure 8.7.2: Diffraction



On the right diagram in the figure there are two barriers with a small opening between them. The opening leaves only the small portion of the wavefront that can pass between the barriers. If the opening is small enough, on the order of one wavelength, then there will not be sufficient interference between the neighboring wavelets for the plane wave to continue propagation forward. The wave will bend around the corners of both barriers, resulting in a complete spherical wave propagating onward from the opening. Thus, the opening acts as a *new source* that generates a spherical wave. This phenomenon is known as *diffraction*.

We will mostly focus on diffraction of light. Since light typically has very short wavelengths, diffraction of light can be observed only when the opening is extremely small. Diffraction of water and sound waves can be more easier observed since their wavelengths are much longer than that of light, thus the openings do not need to be microscopically small to witness diffraction. You can often observe water waves diffract around boats or peers. A sound can appear to be heard from around a corner or a door opening due to diffraction, even though the original source is much further from the door.

Double-Slit Experiment

The *double-slit experiment* is a famous experiment which demonstrates light diffraction through two small slits (openings). Since in this experiment there are now two openings through which the incoming wavefronts encounter, there are two nearby light sources that are generated at each slit according to Huygens' Principle. These are two nearby sources interfere and generate an interference pattern. An animation of the double-slit experiment is shown in the figure below.

Figure 8.7.3: Double-Slit Experiment

The animation above depicts the interference between the two light sources using the wavefront representation. The red wavefronts are coming from the lower slit, while the blue ones are coming from the upper slit. Solid wavefronts are crests and dashed ones are troughs. We see that locations of constructive interference are where two crests or two troughs overlap, while destructive interference occurs when a crest from one slit overlaps with a trough from the other. If you were to place a screen on the right side of the two slits, then the locations of constructive inference will appear as bright spots (equivalent to maximum loudness for sound), and locations of destructive interference would appear at dark spots (equivalent to silence when two sound waves cancel). There will be a central bright spot due to constructive interference in the middle between the two slits, followed by alternating bright and dark regions, also known as *fringes*, on either side of the central bright spot.

You may wonder why we are spending the time analyzing this seemingly particular scenario, but this experiment actually had very important historical significance. The double-slit experiment was first carried out by a British scientist Thomas Young in 1801. As as result it's often referred to as the Young's interference experiment. During that time there was a debate on whether light is a wave or a particle between scientists. Christiaan Huygens was one of the first to mathematically express the "*Wave Theory of Light*" in 1678. Contradictory to this theory, Isaac Newton claimed that light is made up of tiny particles which he described as corpuscles in his *"Corpuscular Theory of Light"* in 1675. Young's experiment demonstrated the wave-nature of light, since only waves can exhibit these interference patterns. This settled the debate for the time being, and the Wave Theory of Light became widely accepted. However, a century later this question was revisited once again with the emergence of quantum mechanics, which showed that light has both wave and particle properties.

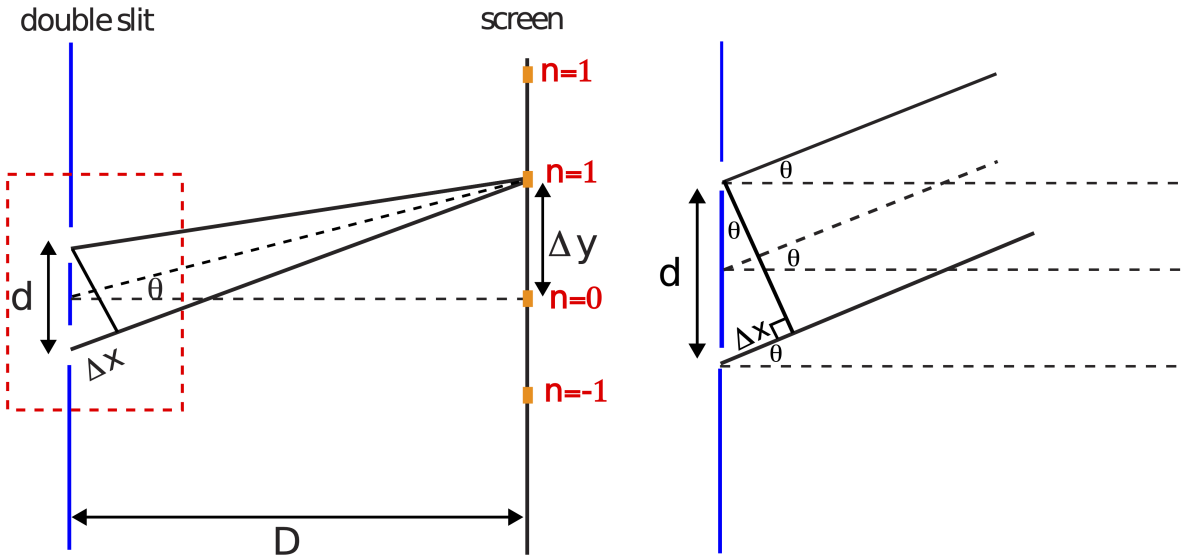
Let us now analyze mathematically the interference of the light sources emerging from the two slits as in appears on a screen placed to the right of the slits. We will assume that the incoming light is *monochromatic* (one color), consisting of one frequency and wavelength, unlike white light which is a combination of a mixture of all wavelengths of visible light. In addition we assume that the light is *coherent,* all the waves generated by the sources are in phase. Most sources of light generate incoherent light with random phases. We need both of these properties to allow to a fixed pattern of interference such that the sources at the slits have the same frequency and phase constant. In modern times, this is typically achieved with *lasers* which generate monochromatic coherent beams of light.

With coherent light entering the slits, we can assume that the two sources generated at two slits are in phase, $\Delta\phi_o = 0$. We are also using monochromatic source of light, so we do not need to consider waves of difference frequencies interfering, as we did for beats in the previous section. Therefore, the total phase difference for a double-slit setup only depends on the path length difference, Δx, which is the difference in the distance from some location on the screen to each slit. The total phase difference is given by:

$$\Delta\Phi = -\frac{2\pi}{\lambda}\Delta x = \begin{cases} 2\pi n: \text{ constructive} \\ 2\pi\left(n+\frac{1}{2}\right): \text{destructive} \end{cases} \quad n = 0, \pm 1, \pm 2, \ldots \tag{8.7.1}$$

The path length difference arises from the rays of light traveling a different distance from each slit to the screen. The figure below will help us visualize the geometry in order to calculate the path length differences for various locations along the screen. The distance from the midpoint between the two slits (the dashed horizontal line in the figure below) to each slit is the same, thus the path length difference is zero, $\Delta x = 0$, resulting in constructive interference for the central bright spot which is labeled n=0 in the figure.

Figure 8.7.4: Geometry of the Double-Slit Experiment



The left side of the figure above illustrates the path length difference to the first bright spot (n=1) above the central one. In this case Δx is not zero, since the ray from the upper slit has a shorter path than the ray from the lower slit to the n=1 location on the screen. The distance marked Δx is the extra distance that the ray from the lower slit needs to travel to the n=1 location.

We want to define the distance Δx in terms of the angle marked θ, between the horizontal dashed line in between the two slits and the line to the location on the screen where interference is measured. The geometry of this situation can be greatly simplified if we assume the two rays are parallel. This assumption is valid if screen is much farther away from the slits than the slit separation $D \gg d$, which is the case in a typical double-slit experimental setup. The right diagram in Figure 8.7.4 shows a "zoomed-in" portion of the rays outlined by the dashed rectangle on the left. If the two rays are parallel all three angles with the horizontal labeled θ are equivalent. With some geometry using the right triangle in the figure, you can then express the path length difference in term of this angle and the distance between slits:

$$\Delta x = d \sin\theta \tag{8.7.2}$$

Plugging the above expression into Δx in Equation 8.7.1, we get the following condition for the bright fringes (constructive interference) and dark spots (destructive interference) on the screen:

$$d \sin\theta = \begin{cases} \lambda n : \text{ constructive} \\ \lambda\left(n + \dfrac{1}{2}\right) : \text{destructive} \end{cases} \quad n = 0, \pm 1, \pm 2, \ldots \tag{8.7.3}$$

The minus sign in Equation 8.7.1 can be absorbed on the right-hand side of the equation, since n is either a positive or a negative interger. The number of bright fringes that appear on the screen has a limit, since the quantity "$\sin\theta$" cannot exceed one. When $\sin\theta = 1$, the angle θ is 90° which means the ray travels parallel to the slits and will never reach the screen. In other words, the number of fringes is determined by the following condition, $n < d/\lambda$. Since n is an integer, if $d \leq \lambda$, there will only the the one central bright spot. When d/λ is a non-integer, then you would need to round down d/λ to determine the number of possible n's.

We can also express the distance between two neighboring bright spots, marked Δy in Figure 8.7.4, in terms of the angle θ and the distance to the screen D:

$$\tan\theta = \frac{\Delta y}{D} \tag{8.7.4}$$

You can then solve the above for θ and plug it back into the θ that appears in Equation 8.7.3 in order to have a relationship between all the distances: slit separation, d, distance to screen, D, and distance between two neighboring bright spots, Δy. However, the resulting equation is cumbersome to work with, since you would need to take the sine of an inverse tangent function. But we can use another useful approximation to greatly simplify the equation. Since we are assuming that $D \gg d$, that implies that the angle is very small, $\theta \ll 1$, which results in the following approximation for both the sine and tangent functions:

$$\sin\theta \backsimeq \theta; \tan\theta \backsimeq \theta; \text{ when } \theta \ll 1 \tag{8.7.5}$$

Using this approximation in Equation 8.7.4 we get $\theta \sim \Delta y / D$. Applying the approximation again in Equation 8.7.3 to solve for the angle to the n=1 bright spot we get $\theta \sim \lambda / d$. Finally, setting the two approximations equal to each other we can solve for the separation Δy between two neighboring bright spots in terms of slit seperation, distance to the screen, and the wavelength of light:

$$\Delta y = \frac{D\lambda}{d} \tag{8.7.6}$$

The distance from the central bright fringe to the first dark spot will be half the distance calculated from Equation 8.7.6, since the dark spot is exactly halfway between two bright fringes. You can also get this result directly by using the destructive Interference condition from Equation 8.7.3. The distance from the central fringe to the second bright spots ($n = \pm 2$) is double the distance in Equation 8.7.6, and so on. Of course, you need to consider the limit of the number of bright spots discussed above and explored further in the example below.

Example 8.7.2

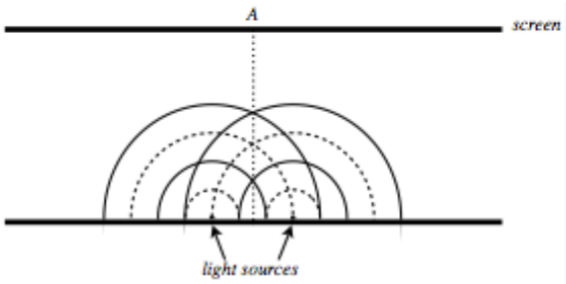
Below are four depictions of two point sources of light (not necessarily caused by two slits), using the wavefront representation. These depictions are snapshots, meaning they are frozen at an instant in time, but the questions below pertain to what happens in real time. Solid lines represent crests, and the dotted lines troughs. For each case, determine the following, and provide explanations:

a) Will these sources create a fixed interference pattern on the distant screen?

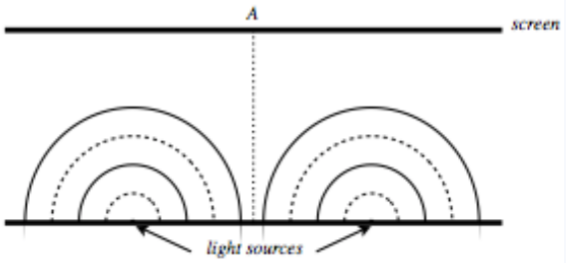
b) If there is an interference pattern, what will appear at the point A on the screen, which is directly across from the midway point between the two sources? That is, will it be a bright fringe, a dark fringe, or something in-between?

c) If there is an interference pattern, how many bright fringes will appear on the screen?

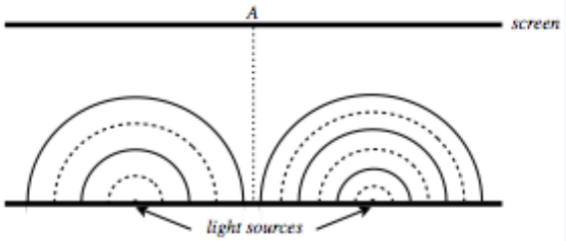
I.



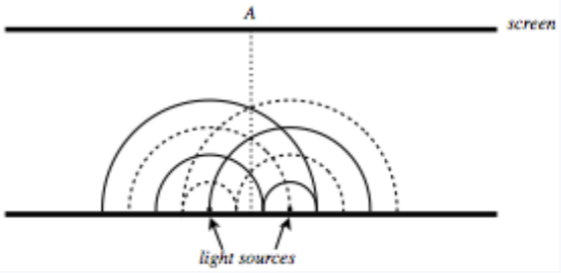
II.



III.



IV.



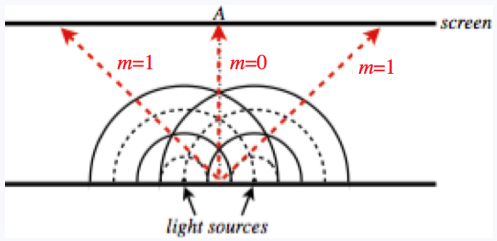
Solution

I.

a) Yes. The sources have the same wavelength (and therefore the same frequency), which means that their interference pattern will not have a time-dependent element to them (i.e. they will not provide the light equivalent of "beats").

b) Bright fringe. The two waves start in phase, and travel equal distances from the sources to get to the center line, so they end up in phase, resulting in constructive interference.

c) One can see by drawing lines through the crossings of crests & troughs that only 3 such lines will strike the screen (parallel to the screen crests match with troughs, so those will not give bright fringes):



We can do this mathematically by noting that these waves start in phase, which means this is equivalent using $d\sin\theta = m\lambda$ for bright fringes, and by noting from the diagram that the two slits are separated by a distance of 1.5λ The fact that $\sin\theta$ can never be greater than 1 puts a limit on m. This is an integer that can't be greater than 1.5, so its maximum value is 1, leaving us with 3 bright fringes.

II.

a) Yes. The same reasons as given above for (I.a) apply.

b) Bright fringe. Same reasoning as II.b

c) Now it is not possible (or at least exceedingly difficult) to draw in the lines that lead to constructive interference, so the mathematical method is the only practical approach. This time the slit separation d is clearly more than 4λ and less than 5λ. This means that the highest integer value of m is 4. With 4 bright fringes on each side of the central bright fringe, the total number is 9.

III.

a) No! These two waves have different wavelengths, and therefore different frequencies, which means that when they interfere, the resulting wave's amplitude (and therefore the brightness) will be time-dependent.

b) N/A

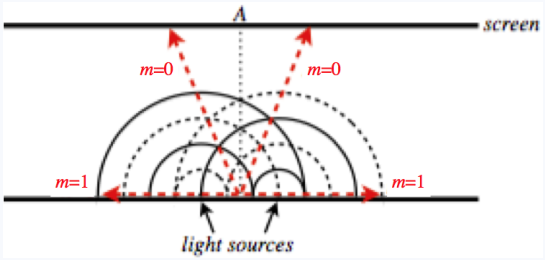
c) N/A

IV.

a) Yes. Back to equal wavelengths.

b) Dark fringe. These waves start out-of-phase by π radians, so when they travel equal distances, they remain out-of-phase.

c) We can once again draw the lines that follow the paths of constructive interference:

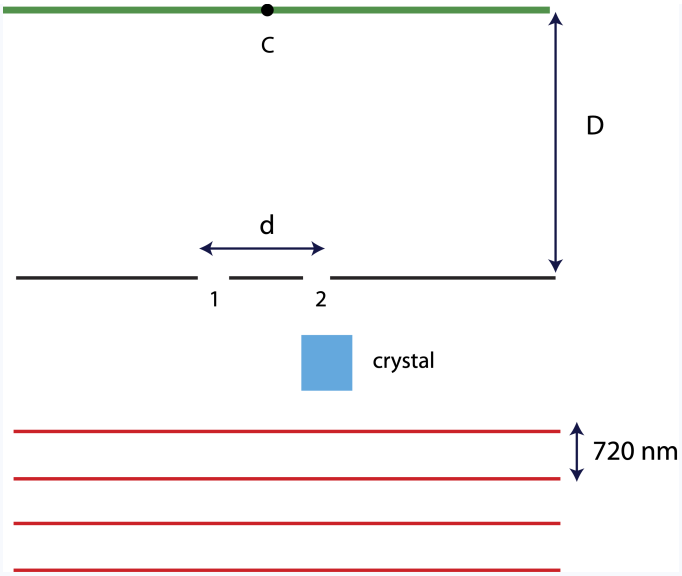


The light sources are separated by 1.5λ as they were once before, but now the condition for constructive interference is different, to make up for the starting phase difference. It is now: $d\sin\theta = (m+1/2)\,\lambda$. We see that there are now two bright spots associated with $m = 0$, and although there is a solution for $m = 1$, it gives $\theta = \frac{\pi}{2}$, which means the light never reaches the screen, so the number of bright spots on the screen is 2.

Example 8.7.3

In a double-slit experiment a red-light laser points toward a double-slit with separation d. The distance between two crests of the plane wave generated by the laser is 720 nm. A crystal is placed in front of slit 2 (as shown in the diagram), such that a wavefront arrives at slit 2 exactly $3 \times 10^{-16} s$ after arriving at slit 1. The interference pattern is measure on a screen which is a distance D from the double-slit. Location C on the screen marks the midpoint between the two slits.

Find the **two shortest** path length differences which will result in constructive interference on the screen. Determine the location of each of these two bright spots on the screen: at point C, to the right of point C, or to the left of point C?



Solution

Since light travels slower inside the crystal, the wavefronts will be no longer reach the two slits at the same time. This will result in a phase constant difference between the two slits. In order to determine $\Delta\phi_o$, we need to determine by what fraction of a period does the crystal delay the wavefront.

The distance between two neighboring wavefronts irepresents the wavelength, $\lambda = 720nm = 7.2\times10^{-7}m$. The period of the red laser is then:

$$T = \frac{\lambda}{c} = \frac{7.2\times10^{-7}m}{3\times10^{8}m/s} = 2.4\times10^{-15}s$$

Thus, the crystal delays the wavefront by the following fraction of one period:

$$\frac{t}{T} = \frac{3\times10^{-16}s}{2.4\times10^{-15}s} = \frac{1}{8}$$

In terms of phase, 1/8 of the cycle delay implies that the wave at slit 1 is $2\pi/8$ ahead of the wave at slit 2, $\phi_1 - \phi_2 = \pi/4$. Solving for the total phase difference which gives constructive interference we get:

$$\Delta\Phi = -\frac{2\pi}{\lambda}(x_1 - x_2) + \frac{\pi}{4} = 2\pi n$$

Note, for consistency it is important to definte $\Delta x = x_1 - x_2$, since the phase difference is defined at $\phi_1 - \phi_2$.

The shortest path length difference, $x_1 - x_2$, corresponds to n=0:

$$x_1 - x_2 = \frac{\lambda}{8} = \frac{720nm}{8} = 90nm$$

Since the sign of $x_1 - x_2$ is positive, the path from x_1 is longer than the path from x_2, implying that this bright spot is to the right of C.

The second shortest $x_1 - x_2$ corresponds to n=1:

$$-\frac{2\pi}{\lambda}(x_1 - x_2) + \frac{\pi}{4} = 2\pi$$

Solving for $x_1 - x_2$ we obtain:

$$x_1 - x_2 = -\frac{7}{8}\lambda = -\frac{7}{8}\times720nm = -630nm$$

Since the sign of $x_1 - x_2$ is negative, the path from x_2 is longer than the path from x_1, implying that this bright spot is to the left of C.

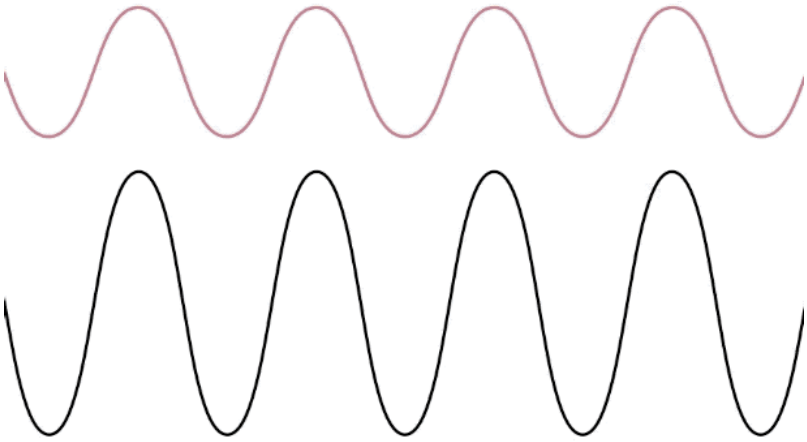
8.8: Standing Waves

Standing Waves Definition

Another important result of wave interference are *standing waves*. Standing waves are formed when a wave encounters a boundary between two different mediums which allows the wave to reflect. Although one source generated this wave, we now have two traveling waves, one outgoing and one reflected. These two waves will interfere in the same manner as do two waves emerging from two separate sources. However, as we will describe in this section, the unique feature of the combined wave in this scenario, is that it no longer has a distinct direction in which it travels. In other words, although there is time-dependent displacement from equilibrium, the positions of the maximum and minima remain in place, thus, the name "standing wave".

The figure below shows an animation of two counter-propagating waves (blue and red on the top panel) which combine to yield a standing wave (lower wave). The two waves have the same frequency since they originate from the same source, and thus the same wavelength since they travel in the same medium. It it clear from the combined wave below why it is called a "standing wave", since we can no longer assign a direction (right or left) to the motion of the wave. Instead, there are specific locations which are fixed in time, where the combined wave exhibits constructive interference, and other locations where the interference in destructive. The amplitude of the standing waves is double the amplitude of the or the original wave. The frequency with which the standing wave oscillates is the same as the frequency of the source waves. The example depicted below is of a one-dimensional standing wave. Standing waves can also be formed in high-dimensions, but the mathematics become much more complex.

Figure 8.8.1: Standing Waves



The location of the standing waves where the amplitude is always at equilibrium are called *nodes*. These are all the locations where the two counter-propagating waves interfere destructively. The distance between two neighboring nodes is half of a wavelength. The location where the wave oscillates away from equilibrium with double the amplitude of the original waves are known at *antinodes*, where the two waves interfere constructively. Antinodes are also separated by half of a wavelength, and the locations of nodes and antinodes alternate.

Digression: Standing Waves Mathematics

Standing waves result from a right-moving wave interfering with a left-moving one. Recall the wave equation for a right-moving wave is:

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

And the equation for the left-moving wave is:

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

One important feature of wave reflections is the change of phase upon reflection. If you watch the animation in the figure above, you can see that at the two ends the outgoing and reflected waves are out of phase. In other words, when the incoming wave hits the fixed end boundary its phase flips by π upon reflection. This results in destructive interference between the outgoing and

reflected waves at the location of the fixed ends. This change in phase is intuitive since the fixed ends are held in place and thus are unable to oscillate, making them natural nodes. On the other hand, for free ends there is no change of phase upon reflection, and as we will see later in this section, antinodes are formed at free ends. For now, we will focus our derivation for a system where both ends are fixed.

The difference in phase between the two waves can be written as:

$$\phi_1 - \phi_2 = \pi$$

If we define one fixed end as the origin $x = 0$, we can define $t = 0$ when $\phi_2 = 0$ which will result in $\phi_1 = \pi$. Plugging in this result and adding the two equations, we obtain the following equation for the combined wave:

$$y_{tot}(x,t) = y_R(x,t) + y_L(x,t) = A\sin\left(\frac{2\pi}{T}t - \frac{2\pi}{\lambda}x + \pi\right) + A\sin\left(\frac{2\pi}{T}t + \frac{2\pi}{\lambda}x\right)$$

We can now use the following two identities to help us simplify the above expression:

$$\sin(x+\pi) = -\sin(x)$$

$$\sin\alpha - \sin\beta = 2\cos\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$

Applying the above trigonometric properties we get:

$$y_{tot}(x,t) = 2A\cos\left(\frac{2\pi}{T}t\right)\sin\left(\frac{2\pi}{\lambda}x\right)$$

You can see from the result above that there the time and the space dependent terms are now separated. The amplitude $2A$ multiplied by the cosine function represents the time-dependent displacement from equilibrium whose maximum displacement is position-dependent, as dictated by the second sine function. A similar expression for the total wave function can also be obtained for other boundaries, such as free ends.

Activity

Log into a free online graphic calculator: https://www.desmos.com/

Enter the following equation that we obtained above for standing waves:

$$y_{tot}(x,t) = 2A\cos\left(\frac{2\pi}{T}t\right)\sin\left(\frac{2\pi}{\lambda}x\right)$$

You will need to select a value for A, T, and λ.

a) Click on "add slider for t" and press play to watch the animation. Do you observe a traveling wave or a standing wave?

b) To change the locations of the nodes and antinodes which parameter do you need to change? Try it out!

c)To change the frequency with which the antinodes oscillate which parameter do you need to change? Try it out!

d) Now plot the two original wave function from which the standing wave originated:

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

$$y_L(x,t) = A\sin\left(\frac{2\pi}{T}t + \frac{2\pi}{\lambda}x\right)$$

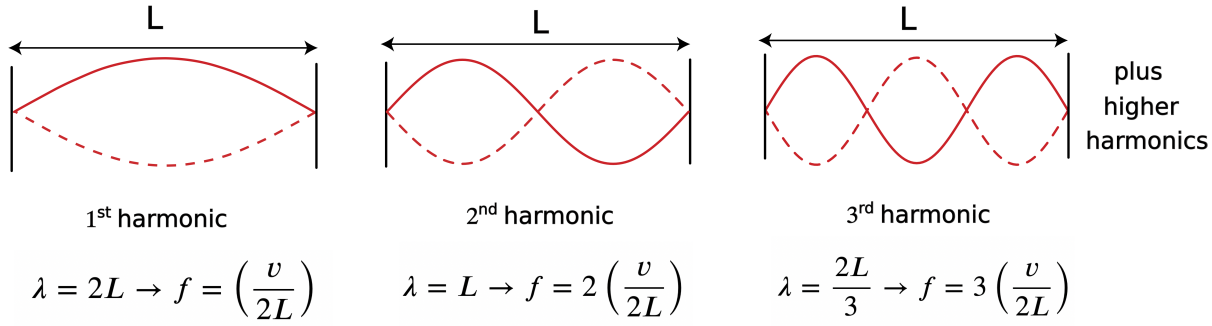
Display all three functions. Instead of running the animation, drag the "t slider" and observe the overlap of the two wave at the nodes and antinodes of the standing waves. From your observations, what type of interference do you see at the nodes and at the antinodes?

Standing Waves Harmonics

When standing waves are formed due to boundaries enforced on a medium through which the waves propagate, such as a string with two ends fixed, we have learned that the two ends have to be nodes. This means that only specific wavelengths can fit on the

string of a fixed length, with nodes always at the ends, and otherwise nodes and antinodes alternating on the length of the string. This is demonstrated in the figure below. The longest possible wavelength that can fit is one where there is one antinode between the nodes. The dashed waved shown represents the motion of the antinode which oscillates from the maximum (solid line) to the minimum (dashed line). The wavelength of this standing wave can be expressed in terms of the length of the string. Since half a cycle is displayed on the string, $\lambda = 2L$, where L is the length of the string when its at equilibrium.

Figure 8.8.2: Harmonics for Two Fixed Ends



The next longest wavelength is shown in the middle panel of the figure, with two antinodes and one node between the two ends. In this case exactly one wavelength fits on the entire string such that, $\lambda = L$. The get the next longest wavelength we add one node and one antinode, such that one and a half wavelengths fit on the string, or $\lambda = 2/3L$. Thus, every time the wavelength is shortened, and extra half of a wavelength fits between the two ends. This gives a general relationship for possible wavelengths on a string of length L:

$$\lambda_n = \frac{2L}{n};\ n = 1, 2, 3, \ldots \tag{8.8.1}$$

where n is a positive integer. Each wavelength then corresponds to a specific frequency. These discrete allowed frequencies are known as *harmonics*. The first three harmonics are shown in Figure 8.8.2. The first harmonic is also often called the *fundamental harmonic*. We can express all the possible frequencies in terms of the frequency of the fundamental harmonic, using the wavelengths in Equation 8.8.1:

$$f_n = \frac{v}{\lambda} = \left(\frac{v}{2L}\right) n = f_1 n;\ n = 1, 2, 3, \ldots \tag{8.8.2}$$

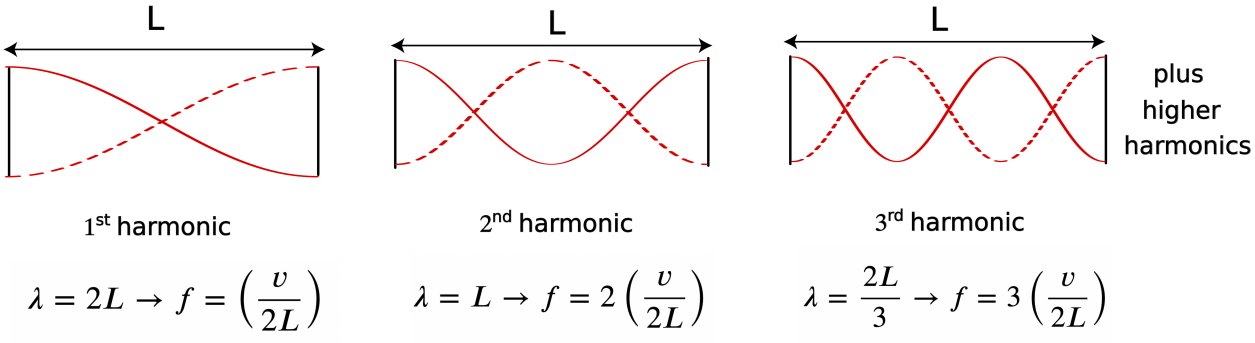
where $f_1 = v/2L$ is the fundamental frequency. Thus, we can think of frequency of n^{th} harmonic, as the n^{th} multiple of the fundamental harmonic.

Alert

For traveling waves we learned that frequency is source dependent and wavelength depends on both the frequency and the medium which determines the speed. These concepts no longer apply for standing waves. In this case, the geometry of the medium determines the allowed wavelengths, which as a result determine the allowed frequencies, which depend on both the wavelength and the medium.

We will now look at all the different types of boundaries that are possible for a wave on a string. Below are the first three harmonics for a standing wave where both ends are free. Free ends results in antinodes at the edges. This situation is not possible for a string which requires tension for waves to propagate, but would work for a rigid system. An example of such a system in shown in Video 8.8.1 below.

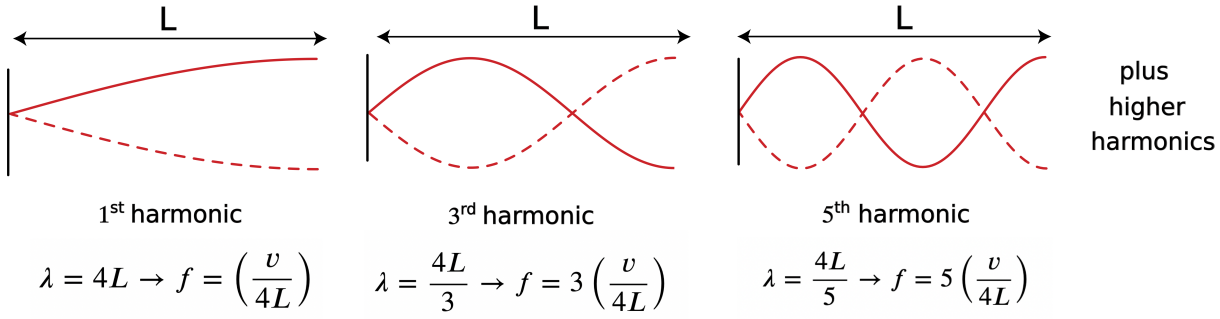
Figure 8.8.3: Harmonics for Two Free Ends



In this case, the longest wavelength has one node between that two antinodes at the ends. The next longest wavelength has one additional node and antinode between the two ends, and so on. Although, the structure of theses standing waves is different from those in Figure 8.8.2, it results in the same relationship between wavelength and the length of the system. Thus, Equation 8.8.2 for the frequencies of the different harmonics applies for this scenario of two free ends.

The last scenario of one fixed and one free end is shown in the figure below. This scenario is also demonstrated in Video 8.8.1. In this case, the longest wavelength corresponds to a node at one end and an antinode at the other one, which is one quarter of a wavelength that fits along the length of the system. Shortening the wavelength with each higher harmonic results in an addition of one node and one antinode. The first three harmonics with their corresponding wavelengths and frequencies are shown below.

Figure 8.8.4: Harmonics for One Fixed One Free End



The general conditions for the possible wavelengths for a system with one fixed and one free end can be written as:

$$\lambda_n = \frac{4L}{n};\ n = 1, 3, 5, \ldots \tag{8.8.3}$$

where in this case, n is a positive odd integer. As for the two previous cases, fixed-fixed and free-free ends, frequency of each higher harmonic can be written as a multiple of the frequency of the fundamental harmonic:

$$f_n = \frac{v}{\lambda} = \left(\frac{v}{4L}\right) n = f_1 n;\ n = 1, 3, 5, \ldots \tag{8.8.4}$$

where $f_1 = v/4L$ is the fundamental frequency for a system with one fixed and one free end. A demonstration of standing waves with various boundary conditions is shown in the video below. The medium used to generate what are called "torsion waves" are multiple rods which are attached together by a central rod allowing the wave to propagate along the system.

Video 8.8.1: System with Two Free Ends and one Free and one Fixed End

In addition to strings, air pressure standing waves can be formed inside cavities or pipes. When air flow is forced through a cavity, reflections at the boundaries of the cavity and the atmosphere creates pressure standing waves, in a similar manner as they are created on strings.

Alert

Recall, for sound waves we express wave displacement in terms of pressure. At an open end of a container (such as a pipe) where a sound standing wave is formed, the pressure is fixed at atmospheric pressure since that end is in contact with the atmosphere. Therefore, the open end of a pipe is a node. At a closed end, the pressure is able to vary as the pressure waves reflected from the hard surface causing periodic compression and rarefactions. As a result, the closed end of a pipe is an antinode. This is an important feature to keep in mind with sound standing waves, since it might be confused with a free end on a rope being a node, while the fixed end as an anti-node. To summarize, for standing waves on strings: fixed end is a node and a free end is an antinode; for sound standing waves inside pipes: open end is a node and a closed end is an antinode.

The sounds from musical instruments are generated due to standing waves formed on strings for string instruments and air standing waves formed inside wind instruments. If you try to vibrate string attached at both ends at a particular frequency, 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. In a standard scenario, such as plucking a guitar string, all the harmonics are excited, with the first or the fundamental being the dominant one. Wind instruments create standing waves in cavities. A flute, for example, is a pipe with both ends open, while a clarinet can be modeled by a pipe with one open end and one closed end, since the mouth covers one end of the instrument.

Example 8.8.1

You are conducting experiments with a torsion wave which is shown in the video above. The video demonstrate how you can have either a free or a fixed end on a torsion wave.

a) You start with both ends fixed. You move one rod at the very edge to generate a pulse, and measure that it takes the wave pulse 0.5 sec to get to the other end of the apparatus. Then you discover that oscillating the rod at a rate of four oscillations per second generates a standing wave pattern. Draw the standing wave pattern that is formed.

b) Now you change the set up such that one end becomes free, while the other end is still fixed in space. You continue to oscillate the rod at the same frequency as before. Explain what happens to the standing wave pattern you observed in a).

Solution

a) For 2 fixed ends the possible wavelengths are the allowed frequencies are:

$$f = \frac{vn}{2L}$$

To find the speed of the wave we use the information given: it takes a wave pulse 0.5 sec to travel the length of the system L, so the speed of the wave in this medium is:

$$v = \frac{\text{distance}}{\text{time}} = \frac{L}{0.5s}$$

To find the standing wave pattern, we need to know the harmonic. The frequency with which you move the rod is 4Hz. Solving for n:

$$n = \frac{2Lf}{v} = \frac{2L \times 4Hz \times 0.5s}{L} = 2 \times 4Hz \times 0.5s = 4$$

Standing wave in the fourth harmonic has 4 antinodes. The diagram is shown below.

b) For one fixed and one free end the possible frequencies are:

$$f = \frac{vn}{4L} = \frac{v}{4L}, \frac{3v}{4L}, \frac{5v}{4L}, \ldots$$

For two fixed ends:

$$f = \frac{vn}{2L} = \frac{vn}{2L} = \frac{v}{2L}, \frac{2v}{2L}, \frac{3v}{2L}, \ldots$$

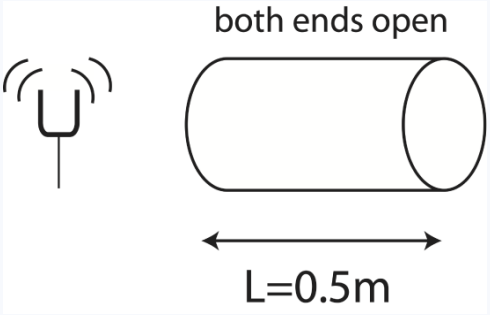
The frequencies can never be the same for for the same length and speed, since the numerators are odd for free-fixed ends and even for fixed-fixed ends (if you make all the denominators 4L). Since 4 Hz formed a standing wave pattern when both ends were fixed, there will be no standing waves when one end is free. Or you can solve directly for n for the when one end is fixed and one is free:

$$n = \frac{4Lf}{v} = 4 \times 4Hz \times 0.5s = 8 \tag{8.8.5}$$

Since n is not an even integer, a standing wave is not possible since by definition the integer must be odd for one fixed and one free end.

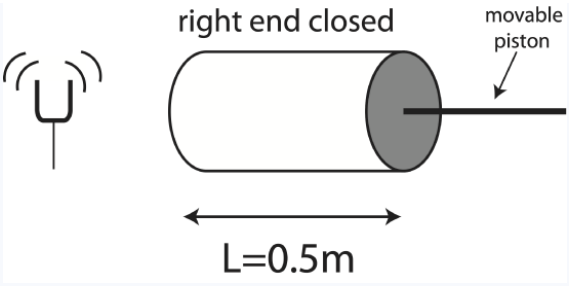
Example 8.8.2

A standing wave in a pipe with both ends open is detected when the sound produced by striking a tuning fork is greatly amplified, a. The set-up is shown below. Let's assume that the length of the pipe is 0.5m, and the speed of sound in air is 340m/s.



a) Different tuning forks of increasing frequency are used until the first resonance (standing wave) is heard in the pipe with both ends open. Then, more tuning forks are used with increasing frequencies until a tuning fork is found such that a second resonance is heard. Plot the standing wave pattern (using displacement in pressure) produced by the final tuning fork. Calculate the frequency of this tuning fork.

b) One of the ends is closed with a piston which is initially positioned on the right side of the pipe, such that the length of the pipe is still 0.5m. Explain what happens with the resonant sound, if the same turning fork as found in a) is used?



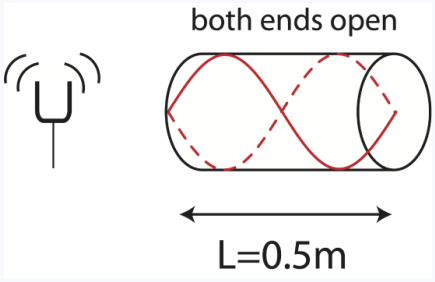
c) The piston shown above is free to move, thus you are able to shorten the length of the pipe. Find the longest length of pipe you can have in order to generate a standing wave pattern using the same tuning fork as in part a). Show the length calculation and plot the resulting standing wave pattern below.

Solution

a) Second resonance implies second harmonic n=2. At both ends of the pipe there must be nodes, since the pipe is open, thus the pressure is atmospheric. Using the equation below for wavelength, and solving for frequency when n=2:

$$f = \frac{vn}{2L} = \frac{2 \times 340m/s}{2 \times 0.5m} = 680Hz$$

The second harmonic is drawn below.



b) A piston placed at the edge of one end changes the system to one open end and one closed end. The open end is a node with the pressure kept fixed by atmospheric pressure, while the other end is an anti-node where the pressure can vary from maximum to minimum as the sound waves collides with the piston resulting in periodic rarefactions and compressions upon reflections.The allowed frequencies a system of a node at one end and an anti-node at the other end are:

$$f = \frac{vn}{4L}$$

where n is an odd integer.

Using the same frequency as in a) and solving for n we get:

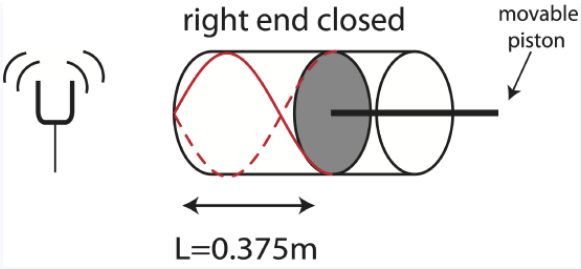
$$n = \frac{4Lf}{v} = \frac{4 \times 0.5m \times 680Hz}{340m/s} = 4$$

However, n must be an odd integer. Thus, no standing wave is possible with this frequency, so no resonance will be heard.

c) Now the length can be adjusted, so we can solve for L such that the odd harmonics are satisfied for a frequency of 680 Hz. Solving the equation in part b) for L:

$$L = \frac{vn}{4f} = \frac{340m/s}{4 \times 680Hz} n = 0.125m, 0.375m, 0.625m, \ldots$$

Since the length of the pipe is 0.5 m, the longest effective length that can be made by shortening the pipe the the piston is 0.375m. This length corresponds to n=3 harmonic which is shown below.



CHAPTER OVERVIEW

9: Quantum Mechanics

9.1: Introduction

We begin our discussion of quantum mechanics with something familiar; try filling a cup with water. We know that we cannot put any more than *one* cup of water in there, right? If someone asked "how much water could go in this cup?" your answer might be “any amount of water less than *one* cup will fit in here.”

Water isn't a continuous quantity though, it's composed of individual H_2O molecules. A standard cup is 250 mL, which is roughly 8.4×10^{24} molecules of water (the exact number is not important. You cannot split a water molecule in half (and still call it water) so you cannot fit *any* amount of water into the cup: you must have zero molecules of water, one 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 *quantized* because the amount water in the cup can only take certain allowed values (multiples of 1 molecule).

In *quantum mechanics*, almost every quantity we encounter (such as energy and angular momentum) can only take certain allowed values. The word “quantum” is derived from the Latin word *quantus*, the same root word as quantity. The name quantum mechanics reminds us that many of the things that we discuss only appear in discrete values. So far, when objects interact by exchanging energy, momentum or angular momentum, we've assumed that we could transfer *any* amount of energy or momentum we wished. However quantum mechanics tells us that we can only transfer these things in discrete pieces (because they are quantized). Here we will focus on quantization of energy and see what the consequences are.

With the example of water, knowledge of molecules makes it easy to visualize why water is quantized. If something is quantized, it's individual pieces are called *quanta* (singular: quantum). The quantum of water is one molecule. Now consider something abstract, like energy. Is the quantum of energy always the same? The answer to this is no: the allowed energies of a system (and therefore the energy quanta), depend on the situation. For example, an atom has different allowed energies than a mass on a spring. One of the mathematically more difficult parts of quantum mechanics is finding the allowed energies of a given system.

We will focus on three specific systems which display rather unique quantization of energy: a particle confined in a one-dimensional box, the hydrogen atom, and a simple harmonic oscillator (a mass on a spring, atomic bonds, etc). We study these three systems to help understand how only being allowed to transfer specific amounts of energy will affect the properties of a system. But first we explore how energy is quantized in the first place, and some of the strange nature of very small objects.

9.2: Particle Model of Light

Photons

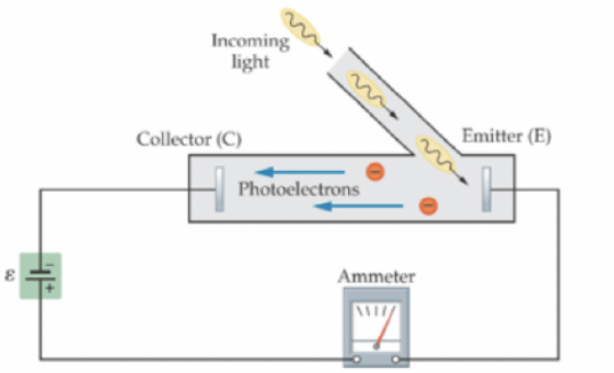
In a previous chapter, we found that light behaves like a wave in a variety of circumstances, such as the interference pattern that light exhibits when it is sent through a double-slit setup. Prior to the two-slit experiments, physicists had been uncertain about the nature of light. Prominent physicists, including Isaac Newton, strongly believed that light was more like a particle than a wave, but the two-slit interference patterns of light could be understood so well with the wave model that for a while the subject was laid to rest.

However, in the early 20^{th} century, several circumstances involving light brought the particle 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.

The Photoelectric Effect

One of the first experiments which is exhibited strong experimental evidence of light behaving like a particle was the *photoelectric effect*, which first led Albert Einstein to develop the particle model of light. In the photoelectric effect, a beam of incoming light shines on a metallic surface. When the beam hits the metal, it eject electrons from the metal and sends the electrons down a tube to a collector. To do so, the light must provide the electrons with enough energy to break their bonds to the metal, and 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. The basic set-up is shown in the figure below.

Figure 9.2.1: The Experimental Setup for the Photoelectric Effect



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 frequency of the light. We can independently vary one or the other, and note the effect, enabling us to determine the appropriate model for this system.

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. If the incident light has less energy than the this binding energy, the electrons remain attached to the plate. In addition, the experiment was able to measure the amount of energy that the ejected electrons had.

Scientists who carried out these experiments observed the following properties:

- Higher intensity beams free more electrons.
- Higher frequency beams result in electrons with higher speeds.
- Changing the beam intensity has no effect on electron speed.
- Changing the frequency of the beam has no effect on the number of electrons freed (provided the frequency is high enough that some electrons are freed).

These results all support the particle model of light for the following reasons. It was postulated that beams with higher intensities contain more quanta of light, known as *photons*. As a result higher intensity beams free more electrons, because more photons are present to transfer energy. However, the amount of energy one photon can transfer to an electron is determined by the photon's frequency. Increasing the frequency of incoming light increases the energy transferred to the electrons, which is why higher frequency beams produce electrons with more kinetic energy. Thus, it was postulated that the energy of an individual photon depends on its frequency the following way:

$$E_{\text{photon}} = hf = \frac{hc}{\lambda} \tag{9.2.1}$$

where $h = 6.636 \times 10^{-34}$ J s is *Planck's constant*, and $c = 3.0 \times 10^{8} m/s$ is the speed of light in a vacuum.Frequency is what determines the type of light we're discussing, whether it different colors in the visible range, a radio wave, or an X-ray. The photons of different colors or different types of light have different frequencies and therefore have different energies. At a particular frequency, one photon is the smallest amount of light that can exist.

To consider the implications of the particle model, it is helpful to think about *monochromatic light*, many photons all with the *same* frequency, like light produced by a laser. We consider two properties of the light: intensity (i.e. brightness) and the amount of energy the light is able to transfer into another system, like an electron orbiting a nucleus. Returning to the photoelectric effect, compare two beams of light with equal intensity, but different frequencies. From our relationship for the energy of a photon in Equation 9.2.1, we conclude that the beam with the higher frequency has photons with higher energy. Thus, the high frequency beam is capable of transferring larger amounts energy into another system. But if the intensities of the beams are same, the total energy transferred by each beam is the same. This tells us that the beam with the higher frequency has fewer photons. But in the wave model, the same intensity of each beam means they must have the same amplitude. The energy in a wave is related to its amplitude, so it would seem both light beams must have equal ability to transfer energy. Clearly, the two models lead to different hypotheses.

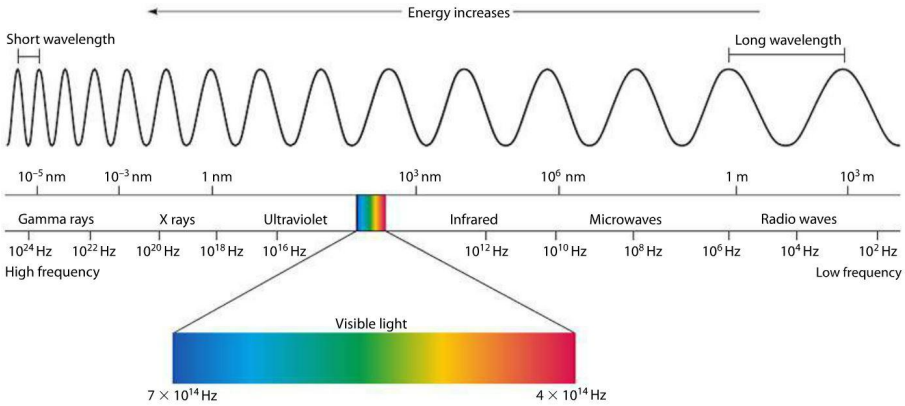
Next, consider the action of increasing the beam's intensity. In the particle model, we would describe this as adding more photons to the beam, but each particular photon still only carries a certain amount of energy. Thus, more electrons are ejected, but each ejected electron still has the same kinetic 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. In the wave model, a greater brightness would indicate a larger amplitude wave; we would conclude that greater intensity waves have the ability to transfer larger amounts of energy into another system. Again, the models make different predictions.

Electromagnetic Spectrum

We just concluded that each photon is an indivisible quantum of light, but photons of different frequencies contain different amounts of energy. So the smallest amount of energy in light is different for each type of light. The light that we experience every day is made up of many photons in a range of frequencies, so we don’t notice the quantized nature of light any more than we notice the individual atoms in everyday materials.

In fact, we categorize light of different frequencies into groups due to its various properties that arise. Below is an image of the *electromagnetic spectrum* (the meaning of this term will become more useful in later chapters). As we will explore in greater detail in later chapters, the sources of electromagnetic waves (i.e. light) are oscillating electric charges. The frequency with which charges oscillate up and down sets the frequency of the electromagnetic waves produced, like similar to how the frequency of the wave on a string is set by a person at the end of the string, oscillating it.

Figure 9.2.2: Electromagnetic Spectrum of Light



For electromagnetic waves, like all the other waves we have studied, the frequency is determined by the source. This seems a little odd, however. Most of us 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". Except the light created in these circumstances is 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 its frequency is closest to red in visible light. While people cannot see this light, some animals can and we can make cameras that can detect this light. For example night-vision goggles operate by detecting infrared light as shown in Figure 9.2.2. If we want an object to give off a significant amount of visible light, we must make its atoms vibrate faster. As we learned in 7A, one way to do this is to increase the temperature. A wood fire, for example, burns at around 1500K. While most of the light is let off in the infrared, enough is let off in the visible range such 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 gives us further insights into the origin of the universe.

Different frequencies of electromagnetic waves can be used for very different purposes. At the lowest frequencies, or at the longest wavelengths, we have *radio waves* as seen in Figure 9.2.2. This is a broad range of frequencies lower than roughly 10^9 Hz. Most of our television and radio programs are broadcast at this frequency. At higher frequencies, there are *microwaves*, which are used in RADAR and microwave ovens. Above 10^{11} Hz we have the infrared, which is the light given off most strongly by objects in the temperature between 3 K and 5000 K. The “narrow” range between 4×10^{14} Hz and 7.5×10^{14} Hz corresponds the spectrum of visible light. In this range, different frequencies correspond to different colors of light that we can see, red in the lower frequency (longer wavelength) and blue in the higher frequency (shorter wavelength) end of the visible range spectrum.

At even higher frequencies we find *ultraviolet light* (UV). This region covers frequencies from 8.6×10^{14} to 3.75×10^{16} Hz and is further broken down into the categories UVA, UVB, UVC, Far UV, and Extreme UV. UV light can be damaging to the skin (UV light given off by the sun is the cause of sunburns). The risk of damage increases with the frequency of the UV light. The sun emits 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 X-ray to describe 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 where extra clarity is needed. 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 refer to frequencies higher than 10^{22} Hz. Because of the quantized nature of light, when discussing gamma rays it's usually more convenient to talk about individual photons than to talk about a continuous wave. For this reason the term "gamma particles" is sometimes used interchangeably with "gamma rays".

Table 1 gives typical approximations for the order of magnitudes of wavelengths from each part of the spectrum. Because wavelength depends on medium, note that he wavelengths presented here are only valid in a vacuum.

Table 1: Order of magnitudes of Spectral Regions

Piece of the Spectrum	Typical Wavelength Size in Vacuum
Shortwave Radio	$\lambda \sim$ Buildings
AM Radio, FM Radio, TV Broadcast	$\lambda \sim$ People
Microwaves	$\lambda \sim$ Insects
Infrared	$\lambda \sim$ Fleas
Visible	$\lambda \sim$ Cells
Ultraviolet	$\lambda \sim$ Molecules
X-rays	$\lambda \sim$ Atoms
γ-rays	$\lambda \sim$ Atomic Nuclei

Table 2 below gives the frequencies and wavelengths of different sections of the visible light spectrum (different colors). The wavelengths presented here are only correct in vacuum, but the frequencies are correct in any medium (because frequency is set by the source of the light).

Table 2: Spectrum of Visible Light

Color	λ range (nm)	λ midpoint (nm)	Frequency (10^{14} Hz)
Red	620-750	700	4.3
Orange	590-620	600	5.0
Yellow	570-590	580	5.1
Green	495-570	540	5.5
Blue	450-495	470	6.4
Violet	380-450	400	7.5

Fire and Chemistry

We remarked that a typical wood fire has a temperature around 1500 K. Emitted frequencies of light at this temperature peak in the infrared range, with some light emitted in the red visible range as well. At low temperature, objects that give off light because of thermal energy (e.g. hot metal) glow a dull red. As the temperature increases, the peak of emitted frequencies becomes higher, and some blue light starts to appear. The dull red becomes an orange as higher frequencies become emitted, and eventually the orange becomes white, which contains many frequencies. The range of frequencies emitted also increases as the temperature goes up, so the light always contains some amount of red in it. We do not observe objects so hot that only glow blue – there is always some contamination from the lower (red) part of the spectrum.

When you study chemistry, you will observe that burning different metals (e.g. in a Bunsen burner) produces different colors of flames. If you have not seen this in chemistry you have probably seen it in a fireworks show – various chemicals in the fireworks give off different colors when they oxidize. This seems to contradict our previous statement that hotter objects glow at different frequencies, but retain red in their emitted light. However, the light we mentioned here is not really thermal light. Thermal light is due to the random motion of charges and the equipartition of energy. The light given off by fireworks is different colors because oxidizing pure substances gives off light whose frequency is determined by the energy available to electrons in the metal. The effect is quantum mechanical, not thermal. We discuss this effect more when we talk about photons and the energy spectrum available to a system.

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 physicists have developed a “quantum model” to explain light, which incorporates all the examples we have discussed so far. However, we should refrain from saying that light is actually this quantum stuff, because 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 beyond the scope of this course. Furthermore, we can answer many questions about light by using the particle model or the wave model of light; both of these simpler models correctly capture aspects of light’s behavior. Many books perpetrate confusion by claiming that light is somehow “both a particle and a wave". We have a good quantum model for light (and electrons, and even whole atoms): in some situations we can simplify this 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. Light and other microscopic phenomena often behave in unfamiliar ways completely outside human experience. Even if we cannot shoehorn quantum mechanics into our regular familiar notions of "particles" and "waves," this does not mean quantum mechanics is contradictory, it just means that the microscopic world is highly counter-intuitive.

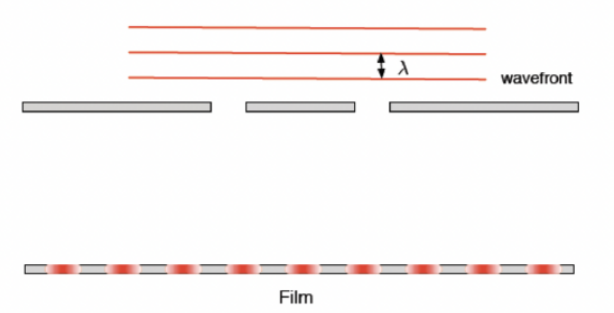
Because the wave-particle “duality” or “contradiction” is bought up so often, it bears repeating. Light can be modeled as particles when it behaves as such and it can be modeled as a wave likewise. We use these models when more complicated behaviors of light can be ignored or simplified, and we recognize that each model has limits and only applies under specific conditions.

9.3: Energy Quantization

Matter Waves

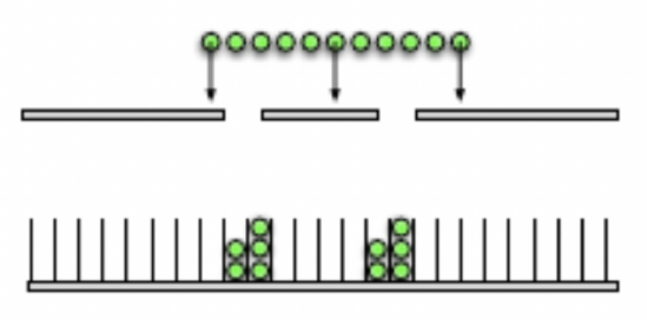
Our definition for a wave has been limited so far, we referred to (material) waves as oscillations of a medium about its equilibrium position in time and space. Light waves light are oscillations in the electromagnetic fields which do no require a medium to propagate. In addition, in the previous section we discussed the validity of a particle model of light. Let us momentarily refer to light as a wave because it obeys the principle of superposition. Superposition gives rise to constructive and destructive interference; an example of this is two-slit interference. Light of a given frequency, going through two narrow slits, superposes to give rise to the bright and dark bands as shown in the figure below.

Figure 9.3.1: Double-Slit Interference of Light



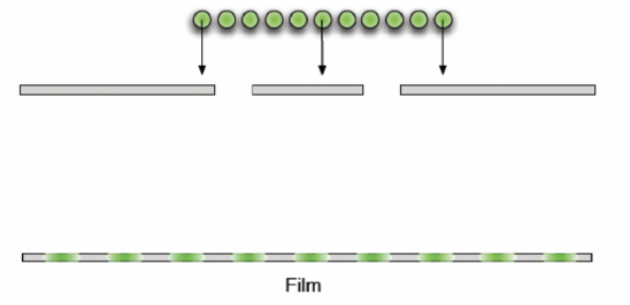
On the other hand, particles cannot interfere like this, they travel through slits one at a time. The figure below shows the pattern particles would make.

Figure 9.3.2: Particle Double-Slit Experiment



What would electrons do if they were also passed through two slits? Since we are used to thinking of electrons as individual particles, we may expect a particle-like interference pattern like the one pictured above, with two bright spots right across the two slits. However, what is actually observed when electrons are sent through a two-slit set up, we get quite an unexpected result – the electrons form an interference pattern similar to waves as pictured below.

Figure 9.3.3: Electron Double-Slit Experiment



In fact, the interference pattern created by the electrons is identical to one created by light with a wavelength given by:

$$\lambda = \frac{h}{p} \approx \frac{h}{mv} \tag{9.3.1}$$

and with the same phase difference between the two slits. The approximation above is valid if $v \ll c$. The wavelength, λ, associated with particle waves is called the *de Broglie wavelength*, named after a French physicist Louis de Broglie who made significant contributions to the development of quantum mechanics. Just as light can exhibit particle-like properties (such as in the photoelectric effect), matter can exhibit wave like effects (such as electron interference patterns).

Particle Energy Levels

Until this unit, our model of energy allowed a particle to have any value of energy. In the quantum mechanics model, this is still true for particles moving freely through space, but the energy of a confined particle is quantized – meaning only certain values of energy are allowed. Like with other models, understanding a few key ideas about quantized 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 vibrating atoms. This is analogies to the allowed frequencies for classical waves. If a wave is generated in a continues medium (such as an infinite string), then you can oscillate it with any frequency to produce the wave. However, when the medium is confined (such as a string tied at both ends), then only a discrete (quantized) number of frequencies of waves can be generated on this medium, as described by standing waves. Similarly, in quantum mechanics when the particle is bounded by some potential (such as an electron bound by the atomic nucleus), then matter standing waves are generated which restrict the allowed frequencies, and thus energies, of the particle.

When we describe the energy of a particle as quantized, 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 can gain or lose. It can only gain the exact amount of energy needed to reach one of the higher energy levels, and it can only lose the exact amount of energy needed to reach a lower energy level. In this case, the particle with 4 Joules of energy can gain either 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 to the particle (unless there were more available energy levels). Similarly, the only lower energy state is 1 J, so if the particle lost energy, it could only lose exactly 3 Joules. The available energy state with the least amount of energy is called the *ground state*. Any higher energy levels are called *excited states*. The next highest state after the ground state is called the first excited state, followed by the second excited state, and so on.

How does a particle change its energy level? If a particle goes to a higher energy level, it must have gained that energy from another system. Likewise, if a particle goes to a lower energy level, that energy must be transferred into another system. We know that energy is conserved if we consider all systems involved in the process of energy transfer. So what systems gain or lose energy to our particles? The most common systems that accept and give energy to systems of small particles are electromagnetic waves (light) and the vibrations of molecules (heat or sound). We will talk about the energy of light in more detail below.

The collection of allowed energies a system may have is called the *energy spectrum* of that system. The levels in the spectrum tell us the total energy the particle is allowed. Typically, the total energy is the sum of the kinetic and potential energies. The energies that a system is allowed to have depend on the potential energy of the system. We call potential energy $PE(r)$ because we typically explore circumstances where PE depends on one variable r. In effect, the potential energy forms a "container" of sorts that confines the particle to a specific range of r. Each unique container has its own set of energy levels. We explore the different spectra of different PE "containers" in the following sections.

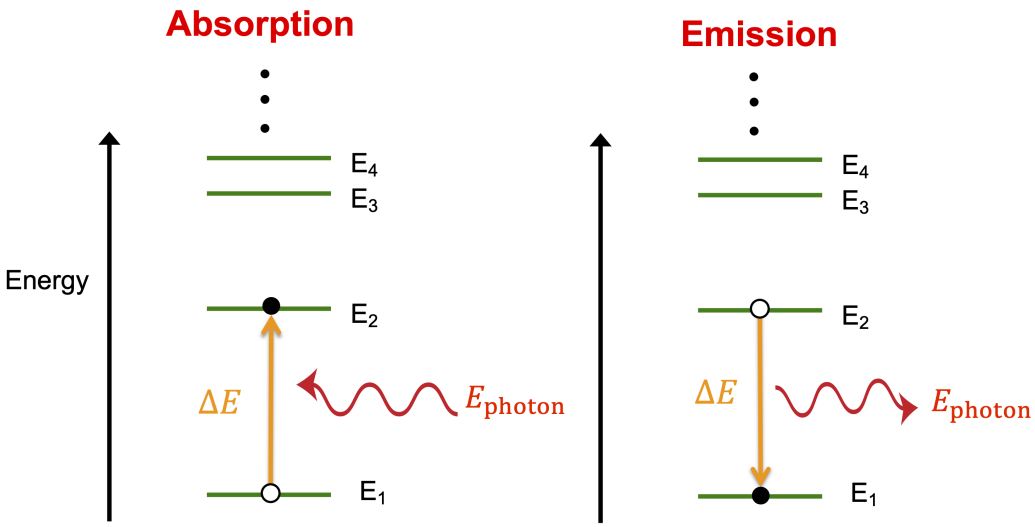
As discussed earlier, the *exact* potential energy of a particle is not important, but the *changes* in potential energy really matter. In quantum mechanics, this is still the case. In effect, we can set "zero" potential energy to be at any level, so in our examples we will establish conventions that make interpreting the results as simple as possible.

Energy Level Transitions

There are multiple ways that a particle whose energy is quantized can change its energy levels, but we will focus on one specific processes depicted below. The figure below depicts arbitrary energy levels for some particle, known as the *energy spectrum*. The horizontal lines represent the quantized energy levels with increasing values upward. The spacing between the lines shows the energy differences between consecutive levels. In this example the energy levels get closer together as the energy increases. In

the following sections we will see examples of different types of energy spectra which is determined by the type of potential by which the particle is confined.

Figure 9.3.4: Energy Level Transitions



A particle can increase its energy by getting the needed energy from a photon. In the process known as *absorption*, a particle absorbs a photon which provides it the extra energy needed to increase to a higher energy level. Since the energy levels are quantized, the energy of the photon has to be exactly equals to the energy difference, ΔE, between two levels. In the figure above the particle is making a transition for its ground state, E_1, to the next higher energy level, E_2. For the transition to be possible the energy of the photon, E_{photon} has to be equals to $E_2 - E_1$. If the photon energy is a little lower or a little higher, the photon will not be absorbed by this particle. The photon's energy is indivisible, so the particle cannot partially absorb the photon. If the incoming photon's energy is equal to the energy difference between the first and the third level, then upon absoption, the particle would transition to the third energy level. Generally, this can be written as:

$$\Delta E = E_{\text{photon}} = hf = \frac{hc}{\lambda} \tag{9.3.2}$$

Since wavelengths of light are often very short, we frequently use the units of nanometer for wavelengths of light: $1nm = 1 \times 10^{-9} m$. This give rise to another useful way to calculate hc in the equation above:

$$hc = 1240 \text{ eV} \cdot \text{nm} \tag{9.3.3}$$

You can use the units above as long as λ is in units of nm. The resulting energy will be in units of *electron volts, eV*. The energy unit "eV" is related to Joules in the following way:

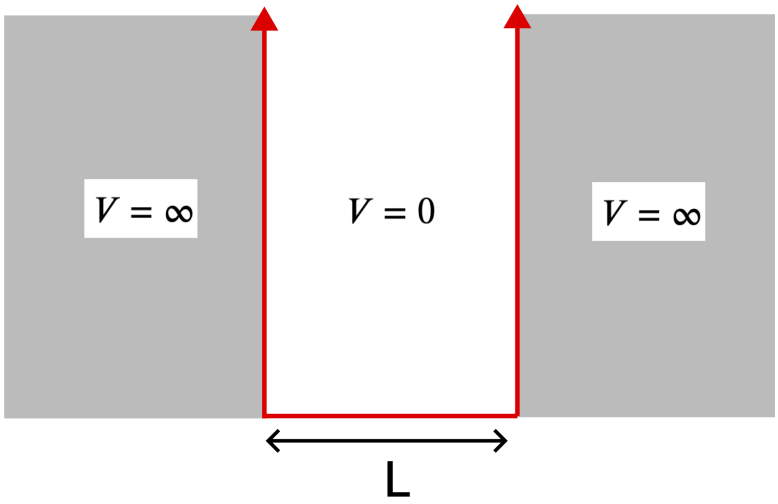
$$1\text{eV} = 1.6 \times 10^{-19} J \tag{9.3.4}$$

Similarly, when a particles drops from a higher to a lower energy level, that change in energy is conserved by a creation of a photon due to the transition. This is known as photom *emission*. This process is depicted above for a particle dropping from a higher energy level, E_2, to a lower energy level, E_1, and as a result emitting a photon whose energy must equal exactly the energy difference between the two levels, $E_2 - E_1 = E_{\text{photon}}$. Even though the final state is the lower energy one, the energy of a photon is always positive, so you still subtract the lower energy from the higher one even when the photon is being emitted and the energy difference between energy levels is negative.

9.4: The Infinite Potential Well

In almost all beginner quantum mechanics classes, the first type of potential which is studied is the "particle in a box". This potential describes a particle which is confined to a physical location, such as a wave confined to the length of a string attached at two ends. Even though the potential is not very realistic, it is the simplest to calculate and still provides a very intuitive understanding of the workings of quantum mechanics. We will describe a particle which is trapped in a one-dimensional box of fixed size but is completely free within 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 this region, hence the name *infinite potential well*. We then set "zero" potential energy to be the energy inside the box. The graph below shows the potential energy of a well with length L.

Figure 9.4.1: The Infinite Potential Well



The infinite well seems to be the least useful of the situations we will study, as very few physical situations are similar to the infinite well. We introduce this system because it has the simplest potential available, zero inside the box and infinite outside of it. 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.

A classical particle trapped in box can have any value of kinetic energy as energy is added to the box. However, we now need to consider the wave nature of our quantum particle. Recall that on a string with two fixed ends, the generated wave had to have zero displacement at the two ends since that part of the string could not vibrate. Analogously, the particle cannot be present at the boundaries of the box, since it would have to have infinite energy. Thus, the displacement of the matter wave, known as the *wave function*, $\Psi(x,t)$, must be zero at the boundaries of the box.

We can apply exactly what we know about standing waves in one-dimension to the particle in a box model, since the two-fixed ends are perfectly analogous to the two walls of the box. We learned in Section 8.8 on standing waves that the allowed wavelengths for a system with two fixed ends are:

$$\lambda_n = \frac{2L}{n};\ n = 1, 2, 3, \ldots \infty \tag{9.4.1}$$

The infinite walls of the potential well assure that the particle cannot exist at the boundary between the wall and the outside. Therefore, the same quantized wavelengths apply to the one-dimensional particle in a box, where L represents the length of the box. Using the de Broglie wavelength equation from Section 9.3 we conclude that if the allowed wavelengths are quantized, so must be the momentum of the particle:

$$\lambda_n = \frac{h}{p_n} \tag{9.4.2}$$

The total energy of the particle is its kinetic energy plus the potential. Given that the potential energy inside the well is zero, the total energy of the particle is:

$$E_n = KE + PE = KE = \frac{1}{2}mv_n^2 \tag{9.4.3}$$

where m is the mass of the particle and v_n is its speed which must be quantized. The kinetic energy can be rewritten in terms of the particle momentum using $p = mv$. Thus, the above equation becomes:

$$E_n = \frac{p_n^2}{2m} \tag{9.4.4}$$

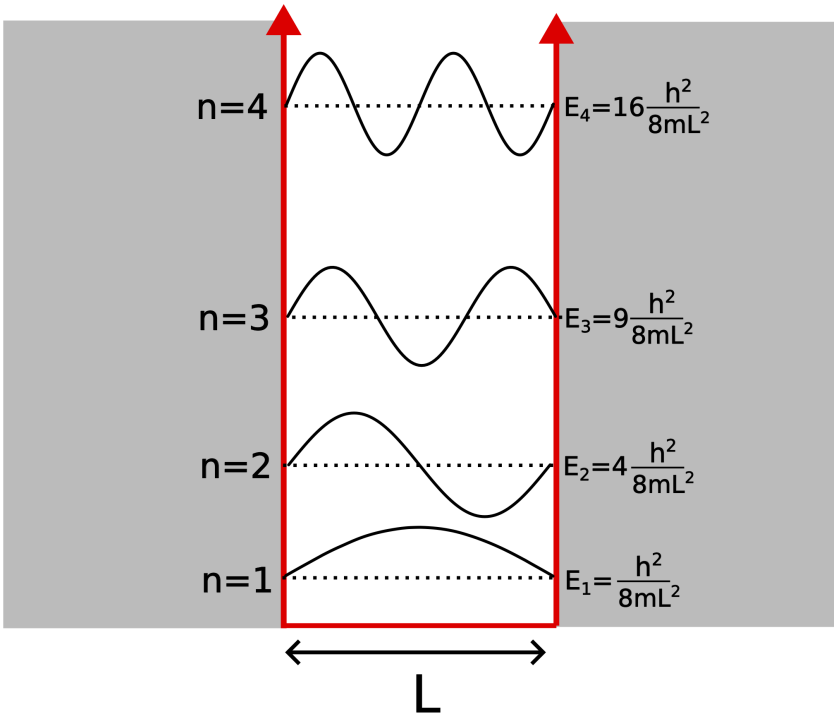
Finally, using equations 9.4.1 and 9.4.2 and plugging into the equation above we get the following expression for the quantization of energy for a particle in a one-dimensional infinite box:

$$E_n = \frac{h^2}{2m\lambda_n^2} = \frac{h^2 n^2}{8mL^2};\ n = 1, 2, 3, \ldots \infty \tag{9.4.5}$$

The minimum amount of energy the particle can have is when n=1: $E_1 = \frac{h^2}{8mL^2}$. The potential energy is zero inside the box, so the particle *always* has some kinetic energy. For a quantum particle in a box it is *impossible* to sit at rest.

The figure below shows the energy spectrum for a particle trapped in a one-dimensional infinite well potential. In addition the wave function corresponding to each energy level is depicted. The wave functions are identical to the standing waves generated on a string with two fixed ends. However, since we are are discussing particles, the wave function has a different meaning. Specially, the wave function is related probability of finding the particle in a specific location in space. Since probabilities cannot be negative, it is the wave function squared, $|\Psi(x,t)|^2$, which gives us the probability directly.

Figure 9.4.2: The Infinite Potential Well Energy Levels



Looking at the wave functions above, the particle has zero probability of being at the boundaries of the wall (nodes) for all energy levels. For the first energy level, the particle has maximum probability of being at the center of the box (antinode). For the second energy level, there are two locations of maximum probability. Of course, the sum of these probabilities cannot exceed one.

For the infinite well potential, the energy levels are proportional to n^2. This means the gaps between lower energy levels are smaller than those between higher energy levels, as depicted above. So as the particle gains energy, it takes *more* energy to transition to a higher level. Also, the energy gap between consecutive levels is smaller if L is bigger. So if the potential well becomes wider, it becomes easier to transition between levels, eventually reaching the classical (continuous energy) limit as L gets very larger or if the particle is very massive. This is why you cannot witness quantum behavior of a macroscopic particle trapped ina macroscopic box. This model applies to atomic particles, such as electron, in a box whose length is on the order of the size of an atom.

Example 9.4.1

An electron trapped in a small box with infinite boundaries is initially in its ground state. It then absorbs a photon which excites it to the first excited state. The wavelength of this photon in λ_o. Then it absorbs another photon which excites it from the first to the second excited state. Express the wavelength of the second photon in terms of λ_o.

Solution

The energy of the absorbed photon has to equal the energy of the transition. When the electron transitioned from the n=1 to the n=2 level the energy of the photon was:

$$E_{\text{photon}} = \frac{hc}{\lambda_o} = E_2 - E_1 = \frac{h^2}{8mL^2}(2^2 - 1^2) = 3\frac{h^2}{8mL^2}$$

For the second transition, from n=2 to n=3, similarly the energy of the photon is given by:

$$E_{\text{photon}} = \frac{hc}{\lambda_1} = E_3 - E_2 = \frac{h^2}{8mL^2}(3^2 - 2^2) = 5\frac{h^2}{8mL^2}$$

where we have defined λ_1 as the wavelength of the photon absorbed in the second transition. Taking the ratio of the two equations above we get:

$$\frac{\lambda_1}{\lambda_o} = \frac{3}{5} \Rightarrow \lambda_1 = \frac{3}{5}\lambda_o$$

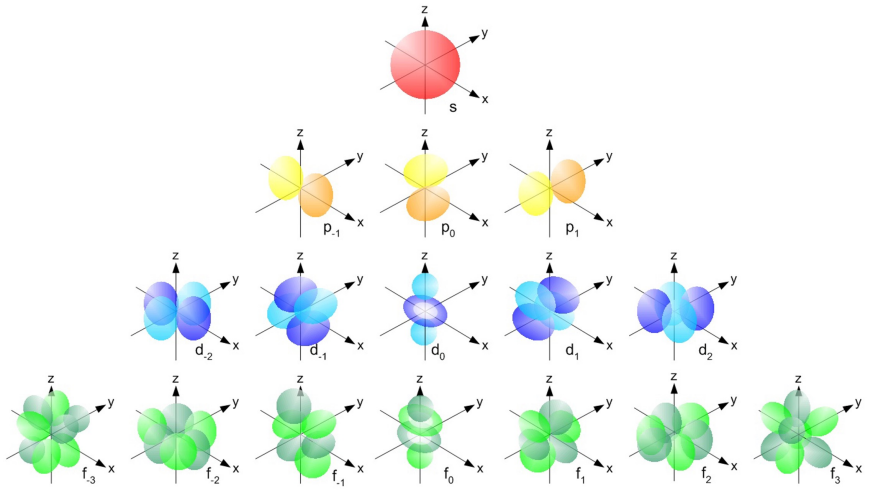
It makes sense that the second wavelength is shorter than the first. The energy levels are getting further apart, thus, higher energy photons with shorter wavelengths are required to make transitions between increasing neighboring energy levels.

9.5: Hydrogen Like Atoms

Hydrogen Atom Energy Levels

The next system we study is a very useful one, an electron bound to an orbit around a nucleus. The simplest case of such a system is a hydrogen atom which has one electron which orbits its atomic nucleus. In this situation, the electron has both kinetic energy and potential energy described by the electrostatic potential which we will study in detail in Chapter 11. Also, the electron orbits the nucleus in three-dimensional space, so the wave functions, described "standing waves" imposed by the electrostatic potential can no longer be described by simple one dimensional waves with alternative nodes and anti-nodes. Instead, these become much more complex, as depicted by atomic orbitals in the figure below.

Figure 9.5.1: Atomic Orbitals



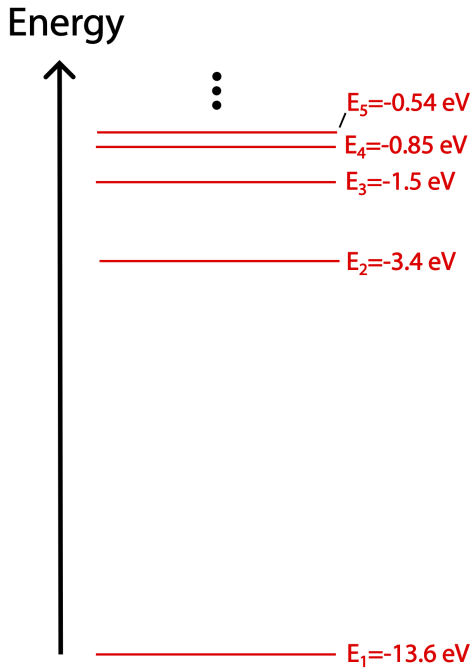
The details behind these orbitals are outside the scope of this class. However, we can still appreciate what they represent. Like for the one-dimensional case, the orbitals represent the probability of finding the electron around the nucleus and correspond to a specific frequency. The colored regions represent anti-nodes, locations where the probability is highest, while the spaces in between are nodes where probability goes to zero.

Quantum mechanics predicts the following expression for the energy levels of a hydrogen-like atom, when both the kinetic energy and the electrostatic potential are incorporated into the total energy:

$$E_n = -\frac{13.6\text{eV}}{n^2}Z^2; \quad n = 1, 2, 3, \ldots, \infty \tag{9.5.1}$$

where Z is the atomic number, or the number of protons. For a hydrogen atom whose atomic numbers is one, (Z=1\), the lowest energy state is $E_o = -13.6\text{eV}$. All of these are one-electron atoms, so atoms with a higher atomic number than hydrogen will be ionized. Helium, $Z = 1$, will have one electron removed, He^+, lithium, $Z = 2$, will have two electrons removed Li^{2+}, and so on. The figure below shows the energy spectrum for hydrogen, displaying the first five energy levels.

Figure 9.5.2: Hydrogen Energy Levels



The energy levels for hydrogen-like atoms get closer together as energy increasing, since the quantized energy is proportional to $1/n^2$. We've established earlier that we could arbitrarily choose "zero" energy to be any amount of energy. Above, we have chosen zero to refer to the energy of an unbound electron, and each energy level shown has a *negative* energy in comparison. 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 infinity). At this point, the electron will be unbound and free, and then will be allowed to have any (positive) value of energy.

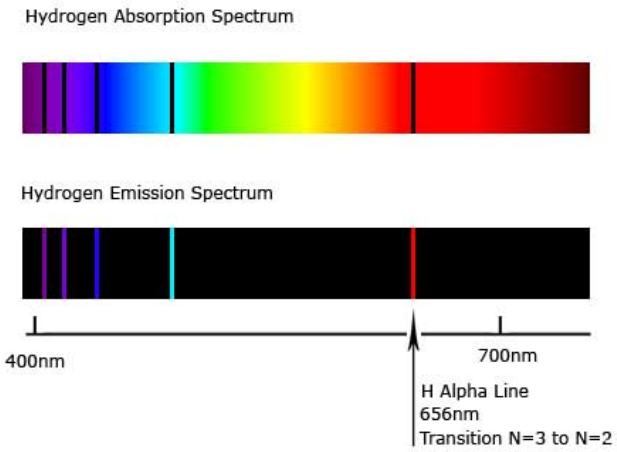
Absorption and Emission Spectra

As before electrons can move up to a higher energy level by absorbing photons, and then fall back down to lower energy levels by emitting photons. Analyzing the frequencies or wavelengths of light absorbed and emitted by different elements is a powerful scientific method of determining the atomic composition of substances, such as distant start in the sky.

When you shine light composed of multiple frequencies on an element, the quantized energy levels will only allow certain frequencies to be absorbed. All other frequencies will be pass through the atoms and can be detected. The *absorption spectrum* of a particular atom will show black lines (absence of light) which correspond to the frequencies of light which were absorbed. The figure below shows the absorption spectrum of hydrogen in the visible range. The red side of the spectrum corresponds to longer wavelengths, thus lower energies, while the blue side shows higher energy transitions. We can see that the levels get closer together as energy increases.

Likewise, the *emission spectrum* of hydrogen shown above, depicts wavelengths of light which are emitted when electrons at higher energy levels fall back to lower ones. You can see in the figure, that the same wavelengths that were absorbed by the electron gained energy were emitted when the electron lost energy. If we heat up a tube of hydrogen, many of the electrons are excited out of their ground states and into higher energy states. As these electrons fall to lower energy levels, they emit photons whose frequencies correspond to the energy transitions. The emission spectrum of hydrogen can be directly calculated from the energy level transitions. The emission spectrum for other elements are more complicated to calculate because other elements have multiple electrons that interact with each other and with the nucleus.

Figure 9.5.2: Absorption and Emission Spectra



Because the atoms of each element have different energy transitions, the emission spectrum and absorption spectrum of each element is unique. This uniqueness is exploited in spectroscopy, where unknown atoms and molecules can by identified by the energies (frequencies) of photons that they 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 chemical fires 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. 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 sodium vapor street lamps, which operate on the principle of exciting sodium atoms, and emit yellow light!

Example 9.5.1

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) How is the detected light different from the incoming one?

b) Considering only transitions that allow the electron to remain bound, determine the longest possible wavelength absorbed.

c) Considering only transitions that allow the electron to remain bound, determine the shortest possible wavelength that could be absorbed.

d) Is the photon wavelength required to make a transition for He^+ longer, shorter, or the same compared for a photon absorbed for the same transition in a hydrogen atom? If different, by how much?

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; it will not contain frequencies corresponding to transition energies in the hydrogen atom. In the case of hydrogen most of the light at high frequencies is absorbed by ionizing atoms.

b) Longer wavelengths of light have lower frequencies. Photons with lower frequencies have lower energies. To find the longest wavelength absorbed we must find the smallest amount of energy absorbed. Since the atoms are close to absolute zero, we can assume that the electrons are initially in the ground state. The lowest energy transition is from the ground state ($n = 1$) to the $n = 2$ level. To make this transition, it must absorb a specific amount of energy:

$$E_{\text{photon}} = \frac{hc}{\lambda} = \Delta E = E_2 - E_1 = -13.6eV\left(\frac{1}{2^2} - \frac{1}{1^2}\right) = 10.2\text{eV}$$

Solving for wavelength:

$$\lambda = \frac{hc}{\Delta E} = \frac{1240\text{eV}\cdot\text{nm}}{10.2\text{eV}} = 122nm$$

c) Shorter wavelengths of light correspond to higher frequencies and higher energies. The highest energy transition available is from $n = 1$ to a very large n, just before the electron is freed from the atom. The initial state is the ground state, with energy −13.6 eV, and the final state approaches 0 eV. For the electron to gain 13.6 eV of energy, it must absorb a photon with that much energy. Mathematically,

$$\Delta E = E_\infty - E_1 = -13.6\text{eV}\left(\frac{1}{\infty} - \frac{1}{1^2}\right) = 13.6\text{eV}$$

Solving for wavelength:

$$\lambda = \frac{hc}{\Delta E} = \frac{1240\text{eV}\cdot\text{nm}}{13.6eV} = 91.2\text{nm}$$

d) For ionized helium the energy levels are four times greater than that of hydrogen since Z=2. Thus, for a transition between any two levels the photon needs to have four times the energy for helium compared to hydrogen. Since wavelength is inversely proportional to energy, the wavelength needs to be four times shorter.

9.6: Quantum Harmonic Oscillator

In Physics 7A we extensively studied the spring-mass system, which is an example of a harmonic oscillator. Classically, a harmonic oscillator describes the motion of a particle whose motion is governed by a restoring force, $F=-kx$, where k is the force constant and x is the displacement from equilibrium. The negative sign in the force equation assures that the force always points toward equilibrium, where the potential energy is minimized. The potential energy of a harmonic oscillator is given by:

$$PE(x)=\frac{1}{2}kx^2 \tag{9.6.1}$$

We saw in Section 8.2 that the harmonic oscillator moves in a sinusoidal manner around equilibrium with frequency:

$$f=\frac{\omega}{2\pi}=\frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{9.6.2}$$

A quantum mechanical analysis of the harmonic oscillator is useful since it can describe similar behavior on a microscopic scale, and it can be a good model for vibrations of molecules in gasses or atoms in solids and can help develop the theory of heat capacity. To solve for the energy levels and wave functions of the *quantum harmonic oscillator* (QHO) one needs to solve the Schrödinger equation with the harmonic oscillator potential energy. Going through the solution is beyond the scope of this course, but we can predict that the energy should be proportional to hf, where the frequency is defined in Equation 9.6.2. If fact it turns out that the energy quantization for a quantum harmonic oscillator is given by the following expression below:

$$E_n=hf\left(n+\frac{1}{2}\right)=\frac{h}{2\pi}\sqrt{\frac{k}{m}}\left(n+\frac{1}{2}\right);\quad n=0,1,2,3\ldots,\infty \tag{9.6.3}$$

Note, that unlike for the two previous systems we studies, the lowest energy state is when n=0. Thus, the energy of the ground state of the system is given by:

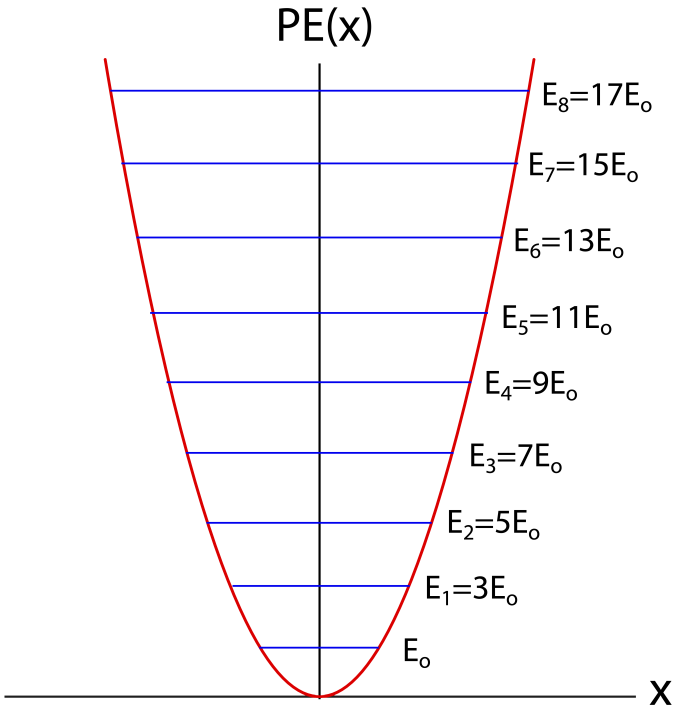
$$E_o=\frac{1}{2}hf \tag{9.6.4}$$

The energies of remaining states can be written in terms of the ground state:

$$E_n=2E_o\left(n+\frac{1}{2}\right)\quad n=0,1,2,3\ldots,\infty \tag{9.6.5}$$

The figure below show the quadratic potential energy of the harmonic oscillator and marks the eight lowest energy levels of a quantum harmonic oscillator. For a classical oscillator any total energy can be added to the system. For example, for a spring-mass, you can add any amount of energy to the system by pulling the spring away from equilibrium by some arbitrary amount before releasing it. Quantum mechanically, however, the figure shows that only discrete levels of total energy are allowed. Another unique feature of the quantum harmonic oscillator is that there are oscillations at the lowest possible energy, E_o. Thus, a quantum harmonic oscillator will vibrate even at zero temperature.

Figure 9.6.1: Energy Levels of a Quantum Harmonic Oscillator



Unlike the case for a one-dimensional infinite box, where the energy level spacing grew with energy, or the hydrogen-like atom, where the energy levels got closer together with increasing energy, for a quantum harmonic oscillator all energy levels are equally spaced. The spacing between two neighboring levels is:

$$E_{n+1} - E_n = 2E_o = hf \tag{9.6.6}$$

The quantization of energy also helps us understand the freezing of vibrational modes that we learned about in Physics 7A. Let us consider a diatomic molecule that vibrates at a frequency f. From Physics 7A, recall that 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 (expressed in Kelvin). For the atoms to vibrate *two* vibrational modes must be activated – one potential and one kinetic. The amount of thermal energy available to two modes is k_BT. To transfer the system to a higher state, it must gain an amount of energy $E = hf$ as derived in Equation 9.6.6. If the thermal energy k_BT available is less than hf, then the molecule does not have enough energy to go up an energy level. We say the vibrational modes are *frozen out* because we cannot transfer energy into them. When the amount of thermal energy is high enough to overcome the gap between energy levels, then energy can transfer into the vibrational energy of the atoms, and we say a vibrational mode has been *activated*. The thermal energy available per mode is controlled by the value of temperature T.

Example 9.6.1

Oxygen gas O_2 has a vibrational frequency of 5×10^{13} Hz .

a) Calculate the temperature required to active the vibrational modes of O_2.

b) Determine the wavelength of the absorbed photon when the vibrational mode is activated.

Solution

a) To calculate the temperature required to activate vibrational modes, we need to set the energy gap of the quantum harmonic oscillator to the thermal energy of vibrational modes for a diatomic molecule:

$$hf = k_BT$$

Solving for temperature:

$$T = \frac{hf}{k_B} = \frac{6.626 \times 10^{-34} J \cdot 5 \times 10^{13} Hz}{1.38 \times 10^{-23} J/K} = 2402 K$$

b) The energy gap is equal to the energy of the photon:

$$E_{\text{photon}} = \frac{hc}{\lambda} = hf$$

The frequency f in the equation above is the vibrational frequency of O_2 and not the frequency of the photon. Solving for the wavelength:

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8 m/s}{5 \times 10^{13} Hz} = 6 \times 10^{-6} m = 6000 nm$$

CHAPTER OVERVIEW

10: Optics

10.1: Introduction

Introduction to Geometric Optics

In Physics 7C so far we have introduced waves and discussed their interactions. Our model of waves has been so useful because it enabled us to apply 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 analyze what happens to waves when they encounter another medium. 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 travel into the next medium, *transmission*. When a wave travels into a new medium the wave will typically bent or deform in some way, a phenomenon we call *refraction*.

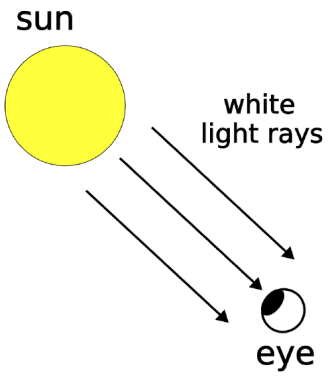
We can combine these effects of reflecting and refracted waves to make the waves appear as if they are being created at different locations than they actually are. If the waves in question are light waves then this means that we will see *images*. While most of our examples will involve light waves, it is important to realize 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 discuss these further.

We will focus of reflection and refraction of light waves, since we are interested in understanding how this phenomena leads to images that we can observe. In Physics 7C we have discussed multiple representations of wave phenomena: the wave equation, wavefronts, and rays. The value of using a given representation depends on the phenomena being studied. As we will see in the following sections, the locations and types of images that are formed because of reflection and refraction depend on the direction of light rays. Thus, the most useful wave phenomena in this chapter will be the direction of light waves, which can be described with ray representation introduced in Section 8.2. It is important to keep in mind that rays are not a physical representing of the light wave itself, such as a laser beam of light. All the light rays represent is the direction of the light wave, connecting the original source to the observer. Light rays will be a very useful geometric tool which will enable us to connect the original light source with the interpretation of that source by an observer. It is possible but much more complex to work directly with wavefronts and Huygen's principle to track the motion of waves as it meets different surfaces, but the light rays give us a very powerful and simple tool to use. The analysis of light propagation using rays is known as *geometric optics*.

How We See Things

Before going too much further into geometric optics it is worth considering how we see things. First, we must establish that eyes only see light rays that fall on them. A light source emits rays going off in all different directions, and we can see the source if some of those rays enter our eyes. Our brain tells us where the light source is by assuming that the light rays travel in straight lines. An good example of a light source is the sun as shown in the figure below. We can see the sun by interpreting a very small fraction of its rays that enters our eyes.

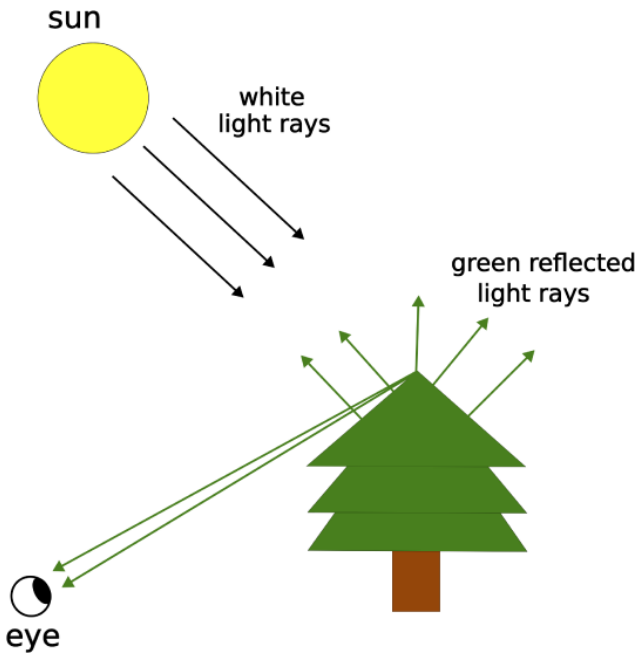
Figure 10.1.1: Direct Light



Another object such as a tree does not give off its own visible light. We know this because we would not be able to see the tree at night time, without any lights on. On a bright day we see the tree because sunrays hit the tree and reflect from it in many directions, some of which reaching our eyes, as pictured below. The light that is reflected is not the same as the incident light, otherwise everything outside would be the same color as the sun. Instead objects reflect certain colors preferentially, and absorb others. The tree leaves, for example, reflect green strongly and absorb most of the other colors. When sunlight (which is a

combination of all the colors) falls on the tree, the green is strongly reflected while colors like red are absorbed. Most of the light that reaches our eyes is green light, this is why the tree appears green. If we can “see” a particular point on the tree, it means some of the rays that reflected from that point enter our eyes. The act of seeing only the top point of the tree is shown below.

Figure 10.1.2: Reflected Light



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 from the top of the tree. 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. If we look at the whole tree, then every point that faces our eye tree that reflects light to our eyes. All points on the tree reflect light in many angles, so our eyes receive light from all points on the tree. The phenomenon of light scattering in all directions when it hits an object is called *diffuse reflection*.

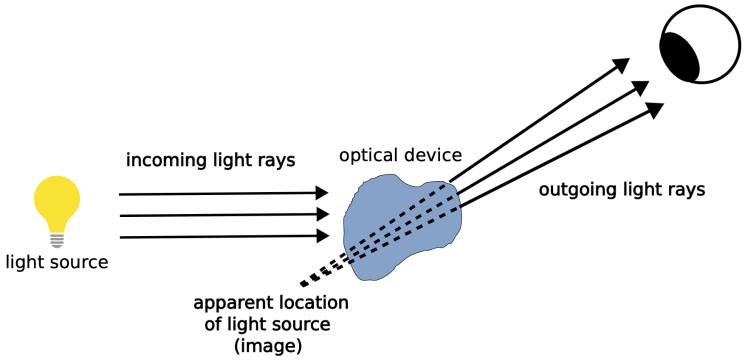
Seeing Images

Since multiple rays enter our eyes from different locations 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 traveled to us in a straight line. This approximation is the what is necessary in geometric optics, when trying to analyze what we see from the original source.

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 the light takes a twisted path to our eyes, then we will judge objects to be at different places, or to be different sizes than they are. 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 the 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*.

This is illustrated in the figure below. Light rays from some original light source encounter some sort of optical device which consists of a medium with reflects and/or transmits light. This results in outgoing light rays with different direction from the incoming light rays. Those outgoing rays are then viewed by an observer, who interprets those rays as coming from a straight now. If you trace back these outgoing light rays, they all meet at a location which is not the location of the original light source. Thus, what the observer sees is the image of the light source and not the light source itself.

Figure 10.1.3: Image Formation



Almost every image used in this section adopts the convention of using solid rays to represent real light rays that actually exists, while using dotted rays to show light rays we interpret to exist. The dotted rays are traced back to the image location, but the solid rays are the actual light rays that we see.

10.2: Reflection

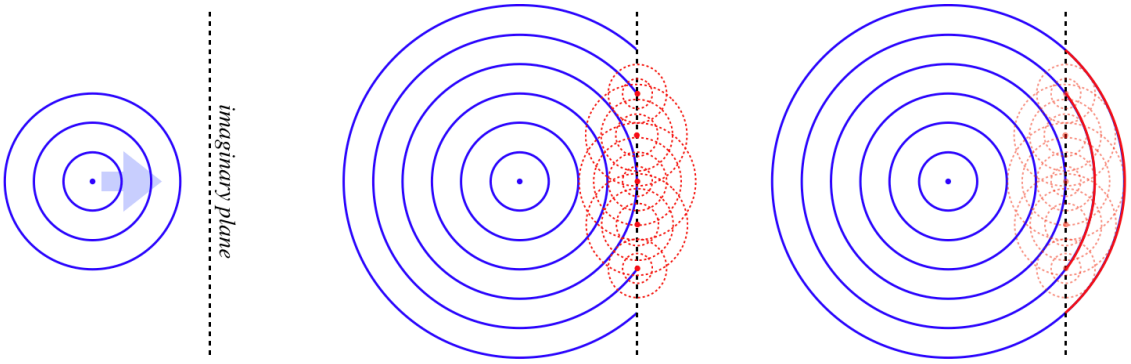
Reflection

When a wave reaches the interface between two different media, typically some of the wave will bounce back into the original medium. This process is known as *reflection*. A familiar example of reflection is optical reflection in mirrors, where light waves reflect off a smooth surface. Another familiar example of reflection comes from water waves, as the waves travel they reflect off objects that are floating in 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. Any kind of wave can undergo reflection.

Our goal is to figure out the direction of the waves upon reflection. If we are observing reflected light knowing the direction of the reflected rays will enable us to extrapolate to the image formed by those reflected rays. In the next section, we will apply these ideas to analyzing images formed by different types of mirrors.

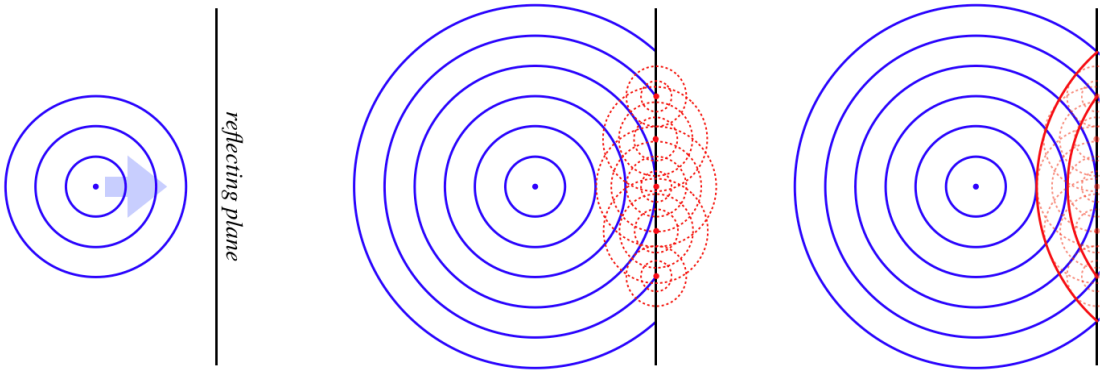
Consider a point source of light that sends out a spherical wave toward an imaginary flat plane, as in the left diagram below. When the wave reaches this plane, then according to Huygens's principle, we can look at every point on the plane and treat it as a point source for an individual wavelet (center diagram below). These wavelets are not in phase, because they are all travel different distances from the source to the plane, and when they are superposed, we know the result is what we see, which is a continued spherical wave (right diagram below).

Figure 10.2.1: Spherical Wave Passes Through an Imaginary Plane



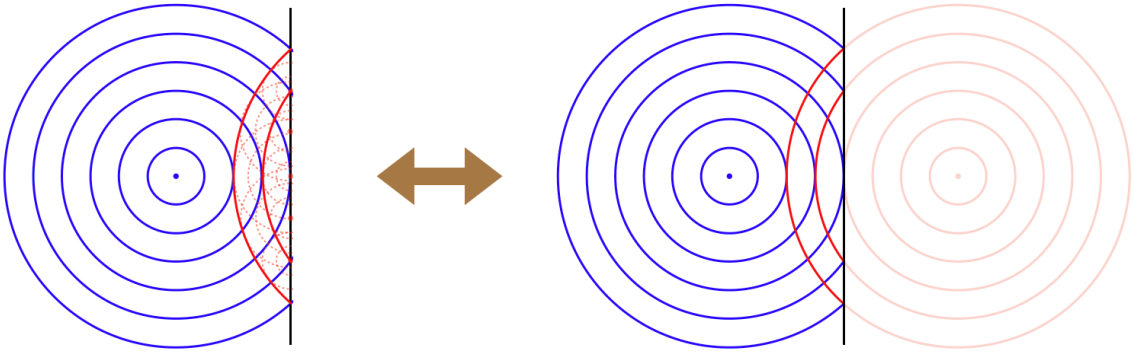
Now suppose the plane is not imaginary, but instead reflects the wave. Every point on this plane becomes a source of a wavelet, but this time, the wave created by these wavelets is going in the opposite direction. The wavelets have the same relative phases as in the previous case, and they are completely symmetric, so they superpose to give the same total wave as before, with the exception that it is a mirror image of the case of the imaginary plane.

Figure 10.2.2: Spherical Wave Reflecting off a Reflecting Plane



Thanks to the symmetry of the situation, it's not difficult to see that the reflected wave is identical to a spherical wave that has originated from a point on the opposite side of the reflecting plane, exactly the same distance from the plane as the source, and along the line that runs through the source perpendicular to the surface.

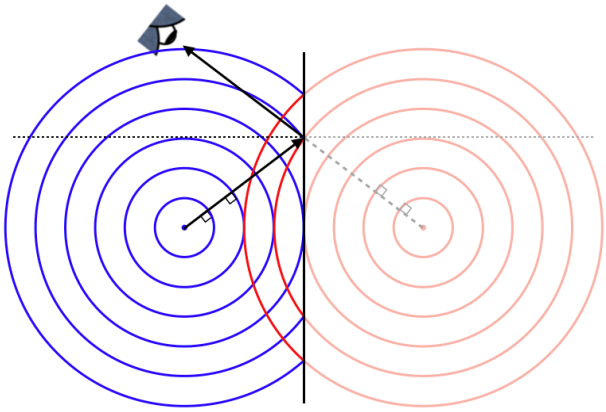
Figure 10.2.3: Image of Reflected Wave



Of course, there isn't *actually* a point light source on the other side of the reflecting plane, it's just that someone looking at the reflected light – no matter where they look from – will see the wave originating from the direction of that point. We call such a point an *image* of the original source of the light.

Now let's put this result in terms of light rays. To do this, we need a source and an observer, and this case, we will require also that a reflection has taken place. Once again drawing the rays perpendicular to the wave fronts, we get the following image.

Figure 10.2.4: Reflection in Terms of Rays

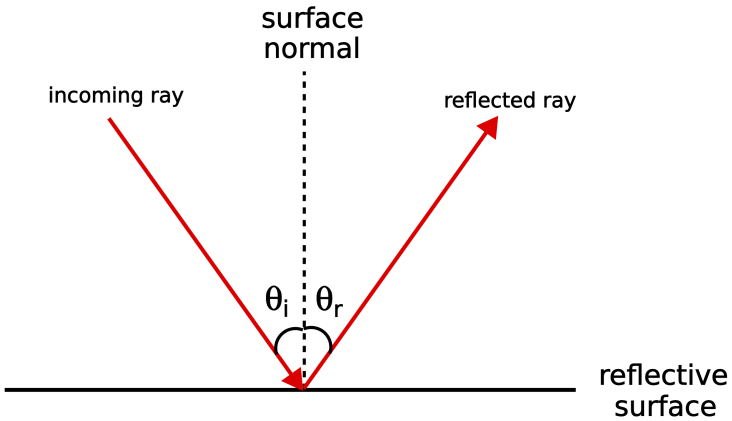


It's clear from the symmetry of the situation that the angle the ray makes with the perpendicular (the horizontal dotted line) to the reflecting plane as it approaches, is the same as the angle it makes after it is reflected. From now on we will stick with rays when describing reflection, without the complication of spherical waves and Huygen's principle which governs how the waves reflect when they encourage a surface.

The Law of Reflection

The above result gives us the *law of reflection*. This law is summarized in the image below.

Figure 10.2.5: Law of Reflection



In optics, we will measure the relevant angles in terms of the angle from the *surface normal*, a line which is perpendicular to the surface from which the ray reflects. The law of reflection states that the incoming angle is defined as the *angle of incidence*, θ_i, is equals the reflected angle, the *angle of reflection*, θ_r:

$$\theta_i = \theta_r \tag{10.2.1}$$

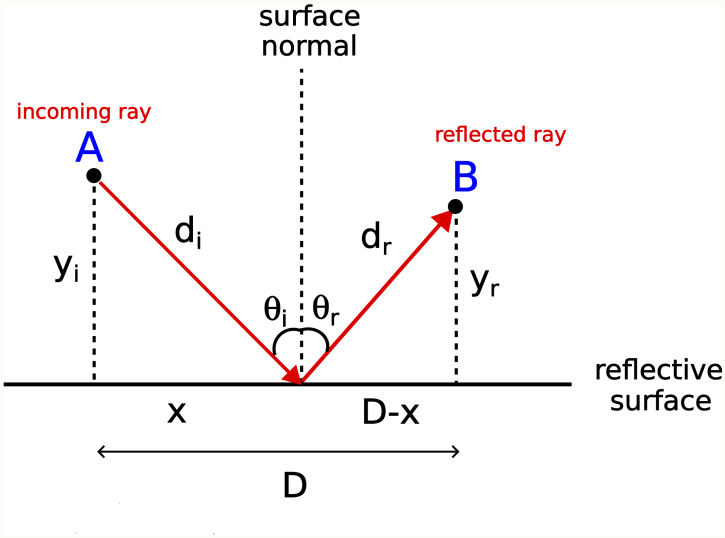
Law of reflection can also be obtained by *Fermat's principle* that states:

"Light travels between two points along the path that requires the least time, as compared to other nearby paths."

This rather simple idea will lead us to the same law of reflection demonstrated by Huygen's Principle above. For those who are interested to see the derivation of the law of reflection from Fermat's principle you can read the "digression" below.

Digression: Law of Reflection from Fermat's Principle

Fermat's principle states that light travels between two points such that it takes the shortest amount of time. Consider the diagram below.



If light was to travel from point A directly to point B, the shortest path would clearly been the straight line connecting the two points. Now let's consider a ray originating from point A that must first reflect from the surface before reaching point B. Our goal is to find the the distance that the ray travels, $d_i + d_r$, such that the time is minimized. Another way to pose this question is to solve for the distance x, marked in the figure, which insures the shortest path between A and B with a reflection.

The time it takes for the ray to travel from point A to point B is the total distance total distance, $d_{\text{tot}} = d_i + d_r$, divided by the speed of light, c:

$$t = \frac{d_{\text{tot}}}{c} = \frac{d_i + d_r}{c}$$

Using Pythagorean theorem for each right triangle formed in the figure, the above equation can be written in terms of the horizontal distance x:

$$t = \frac{\sqrt{x^2 + y_i^2}}{c} + \frac{\sqrt{(D-x)^2 + y_r^2}}{c}$$

The vertical distances y_i and y_r and the horizontal distance D are are fixed for given points A and B. To find the distance x which minimizes time, we need to differentiate time with the respect to x and set it to zero:

$$\frac{dt}{dx} = 0 = \frac{x}{c\sqrt{x^2 + y_i^2}} - \frac{D-x}{c\sqrt{(D-x)^2 + y_r^2}}$$

Expressing the above equation in terms of the angles θ_i and θ_r we get:

$$\frac{dt}{dx} = 0 = \frac{1}{c}(\sin\theta_i - \sin\theta_r)$$

Rearranging, we arrive at the law of reflection:

$$\sin\theta_i = \sin\theta_r \Longrightarrow \theta_i = \theta_r$$

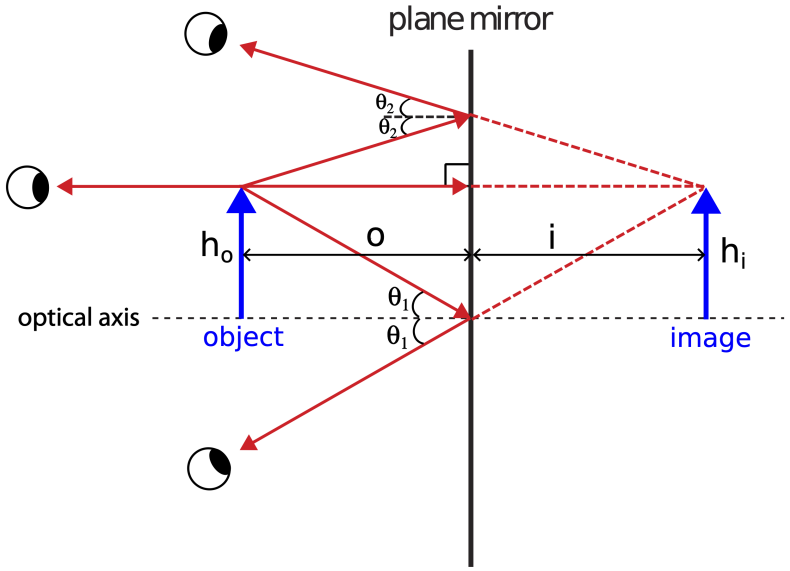
10.3: Mirrors

Plane Mirror

In the previous section we described the law of reflection, and now we will see that this simple law will help us understand how reflective surfaces create images. We will focus on mirrors as the standard reflective surface, although there are many other surfaces such as a clear lake which can produce a sharp reflective images. Let us start with the most standard mirror which we use in our daily lives. This mirror is known as the *plane mirror*, simply due to its flat shape.

Below is an example of an optical setup for a plane mirror depicted as a vertical line with the reflective surface on the left side. A physical *object* is placed in front of the reflective surface of the plane mirror. The horizontal dashed line that is perpendicular to the mirror is known as the *optical axis*, a reference from which we measure the heights of the objects and images formed.

Figure 10.3.1: Image Formation by Plane Mirror



The "object" could be any physical object or a source of light, but we often depict it as an upright arrow. The arrowhead will allow us to distinguish between upright and upside down orientations since, as we will see shortly, some images will have inverted orientations. The object emanates rays in all direction. Some of those rays hit the reflective surface of the mirror and reflect back. An observer standing in front of the mirror will then detect the reflected rays and interpret them as originating from some location from which the rays take a straight path. In other words, the observer detecting the reflected rays does not have any information about the ray initially reflecting before reaching the eyes of the observer.

For convenience, we often choose a few rays originating from the tip of the arrow to analyze. The distinction between the bottom and the top of the arrow is relevant since it will allow us to determine the orientation of the image. In addition, it enables us to find the location of the image of the tip of the object, from which we can extrapolate the image of the remaining object.

In Figure 10.3.1 above we choose three rays and apply the law of reflection to find the path of the reflected rays. The ray which is parallel to the optical axis will meet the mirror perpendicular to its surface (or parallel to the surface normal), which means that it will reflect right back along the same line. To other two rays are shown with the incident angle equal to the reflected angle relative to the surface normal.

In order to determine where the three reflected rays appear to be coming from to the observer, we trace the rays back to determine if they intersect. The location where all three rays meet will be the position of the image of the arrow's tip, since all three rays originated from the tip. As we see from the figure the reflected rays do not cross anywhere in front of the mirror. However, if you continue tracing the rays behind the mirror, you find a specific location where all of them cross. If you drew more rays originating from the tip of the object and applied the law of reflection at the surface of the mirror, you would find that those rays when traced back would also meet at the same location behind the mirror. Since all the rays originating from the object appear to be coming from the location behind the mirror, we observe the *image* of the object behind the mirror. An image is the appearance of the object at a location different from the physical object.

The reason why the lines behind the mirror are drawn with dashes is that the they are no longer physical rays, but simply the extrapolation of the rays to the location behind the mirror. There is no light penetrating the mirror, yet every time we look in a mirror we see ourself as if appearing from behind the mirror. This type of image is known as *virtual*, since it is not *real* light rays that form the image, but rather the tracing of real rays to the location of the image. This definition of a virtual image will become more relevant later, when we compare virtual images to real ones.

The ray that is perpendicular to the mirror and reflects along the same line establishes the fact that the *height of the object*, h_o, is equal to the *height of the image*, h_i. This means that the image is neither *magnified* (enlarged) or *de-magnified* (reduced), but remains the same size as the object. (Below we will encounter other types of mirrors which can indeed create magnified or de-magnified images). The distance from the mirror to the object is known as the *object distance*, o. The distance from the mirror to the location of the image is known as the *image distance*, i. The two right triangles in Figure 10.3.1, one from the optical axis at the mirror to the tip of the object and the other to the tip of the image are identical. From this we can conclude that the object distance is equal to the image distance.

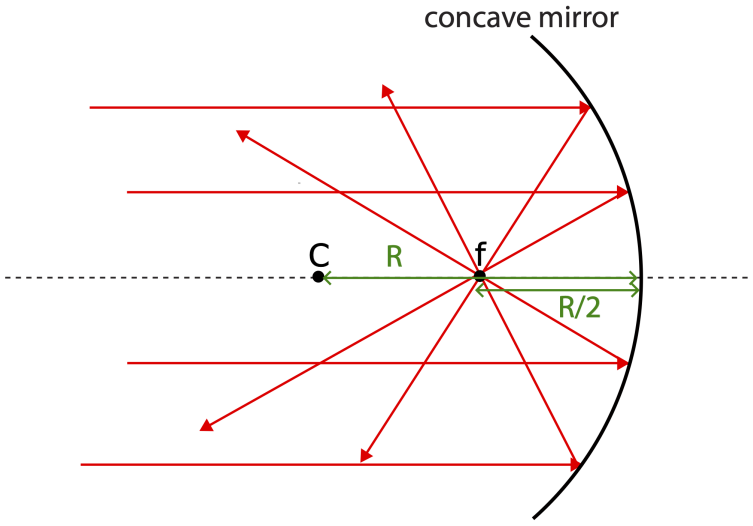
Spherical Mirrors

Although plane mirrors are the most common mirrors we encounter daily, the images they produce are simple to interpret. Much more interesting optical phenomena emerges when we look at mirrors with non-planar shapes. One such group of mirrors is known as *spherical mirrors*, due to their shape being is a section of a sphere.

Concave Mirrors

Below is a diagram of one such spherical mirror, a *concave mirror*, named after its shape. A good way to remember the shape of a concave mirror is to think about a "cave". Since a concave mirror is a section of a sphere, it has a well defined *center of curvature*, C. The distance from the center to the mirror is the *radius of curvature*, R. We draw the optical axis of the spherical mirror to go right through the center. One special feature of a concave mirror is what happens to incoming plane waves upon reflection. Consider rays that are coming parallel to the optical axis, such as rays originating from a distance object or from a laser. If we apply the law of reflection to all parallel incoming rays, we discover that they all *converge* (meet) at one point along the optical axis. This point of convergence is called the *focal point* of the mirror. This result is approximately true if we assume that the incoming rays that are close to the optical axis, due to a small angle approximation. In other words, the simplified model of spherical mirrors that we develop in this chapter applies for objects whose size is much smaller than the radius of curvature of the mirror, or when objects are much closer to the mirror than its radius. We will only consider optical effect using this approximation as it applies to mirrors and lenses (covered in a later section).

Figure 10.3.2: Concave Mirror



The distance from the focal point to the mirror is called the *focal length*, f. We will not go into the details of the proof, but it can be shown using the small angle approximation that the focal length is equal to half the radius of curvature:

$$f = \frac{R}{2} \tag{10.3.1}$$

Our next goal is to determine what kind of images concave mirrors produce of objects placed near the mirror. For the plane mirror, we choose a few rays, used the law of reflection to find the path of the reflected rays, and found where those reflected rays converge in order to find the image. We use a similar procedure for spherical mirrors, except we make use of "convenient" incoming rays which will allow us to immediately determine the direction of the reflected rays, without needing to calculate the angle of incidence and reflection. One such ray that travels from the object parallel to the optical axis will reflect through the focal point, by definition of the focal point. By symmetry of the law of reflection, a ray that goes through the focal point will reflect parallel to the optical axis. Another "special ray" is one that goes directly through the center of the sphere. Using a property that line that originates from the center will be perpendicular to the surface of the sphere, we find that, based on the law of reflection, this ray will reflect straight back along the same line since the angle with the normal is zero.

The three "special" rays described above, known as the *principle rays*, for a concave mirror:

- Principle ray #1: incoming ray parallel to the optical axis will reflect through the focal point.
- Principle ray #2: incoming ray that goes through the focal point will reflect parallel to the optical axis.
- Principle ray #3: incoming rays that goes through the center of curvature will reflect straight back.

These principle rays are depicted with an animation below.

Figure 10.3.3: Principal Rays of a Concave Spherical Mirror

Using rays to determine the location, orientation, and the size of the image is known as *ray tracing*. We can see that the image where three principal rays in the animation intersect is in front of the mirror, closer to the mirror than the object, and below the optical axis. Unlike for a plane mirror in Figure 10.3.1 where the rays had to be traced behind the mirror to find their intersection and thus image location, in this scenario the actual physical rays intersect in front of the mirror, which make this a *real image*, compared to a virtual mirror that a plane mirror created. What make real images distinct from virtual ones, is that real images can be projected on a screen. You can place a screen at the location of the image, allowing those rays to be reflected in all directions, so the image can be seen now seen from multiple angles. You cannot project a virtual image on a screen by placing it behind the mirror, since there are no physical rays present there that can reflect from the screen.

If you were to use the same principle rays coming from the middle of the arrow, you would find that they meet at the same location from the mirror as the rays coming from the tip, but halfway closer to the optical axis. Therefore, since the tip is further from the optical axis compared to the rest of the arrow, the image of the object is *inverted*, appears to be upside down compared to the vertical orientation of the object. The image also appears to be smaller than the object as can be seen in Figure 10.3.3. To determine how much smaller the image compared to the object, you can simply use a ruler and measure the heights of the object and image from the ray tracing in the figure. Likewise, you can measure the distance from the mirror to the image and compare it to the object distance. However, we would like to develop more accurate mathematical relationship between object and image distances and heights. This can be done with pure geometrical arguments.

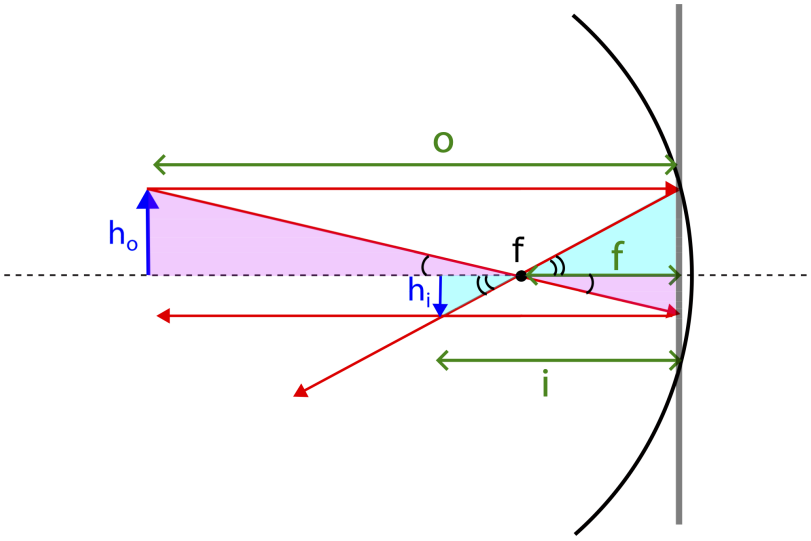
Alert

Although we will focus on the three principle rays coming from the tip of an object, when determining image sizes and positions, it is important to remember that there are an infinite number of rays that are coming from all position of the the object,

many of which will hit the mirror and reflect. All of these rays will then converge at the position of the image for a real image or appear to originate from the location of the image for a virtual image.

The figure below shows a real image formed by a concave spherical mirror. For the purpose of clarity only two principle rays are shown in the figure. The height of the object is labeled as h_o, while the height of the image is marked h_i. The horizontal distance from the object to the mirror is the object distance, o, and the distance from the image to the mirror is the image distance, i.

Figure 10.3.4: Deriving Equations for a Concave Spherical Mirror



Since we are using the small angle approximation (we assume that all distances are close to the optical axis), the mirror can be approximated as flat where light is reflected, as shown by the bold vertical line. This helps us relate these distances using the triangles shown in the figure. The two light pink triangles are *similar*, since they are both right triangles and share the same angle as marked. Using the property of similar triangles we get the following relationship:

$$\frac{h_o}{h_i} = \frac{o-f}{f} \tag{10.3.2}$$

The two turquoise triangles are similar as well. Using the same property of similar triangles we obtain another relationship for the ratio of heights:

$$\frac{h_o}{h_i} = \frac{f}{i-f} \tag{10.3.3}$$

Since the left-hand sides of the two equations above are the same, so we can set the right-hand sides equal to each other:

$$\frac{o-f}{f} = \frac{f}{i-f} \tag{10.3.4}$$

With some algebraic manipulations of the equation above we can to obtain the desired relationship between the focal length, the object distance, and the image distance:

$$\begin{aligned} f^2 &= oi - of - if + f^2 \\ oi &= f(o+i) \\ \frac{1}{f} &= \frac{o+i}{oi} \end{aligned} \tag{10.3.5}$$

Rewriting the above result in fractional form we arrive at the following (small angle) *mirror equation*:

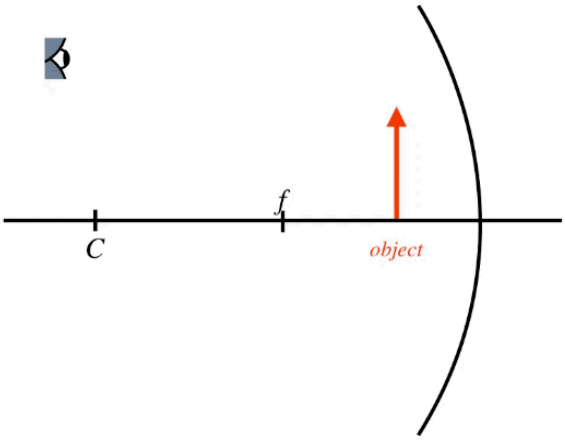
$$\frac{1}{f} = \frac{1}{o} + \frac{1}{i} \tag{10.3.6}$$

Notice that when the object is placed exactly at the focal point, $o = f$, the equation tells us that the image is formed at infinity. This result comes directly from the property of the focal point of a spherical mirror, all parallel rays reflect through the focal point. Due

to the symmetry of the law of reflection, this implies that all rays that originate at the focal point when the object is placed there will be reflected parallel to the optical axis, placing the image at infinity.

Something interesting happens when the object is placed between the mirror and the focal point, $o < f$. Equation 10.3.6 tells us that the image distance becomes negative since $1/f < 1/o$. The animation below shows ray tracing with two principle rays shown for the scenario of an object placed between the mirror and the focal point. Since the object is to the right of the focal point and the center of curvature, the principle rays that would be going through those points to reach the mirror are now the rays that are coming from the direction of the these two points toward the mirror. The red dashed lines on the side of the mirror in the animation are to stress that the rays are lined up with f and C. The last principal ray, one that travels parallel to the optical axis and reflects through the focal point, is not shown in this animation. Thus, the reflection of these rays is dictated by the same rules, the ray lined up with the focal point will reflect parallel to the optical axis, and the one lined up with the center will reflect back along the same line.

Figure 10.3.5: Principal Rays for an Object Close to a Concave Spherical Mirror



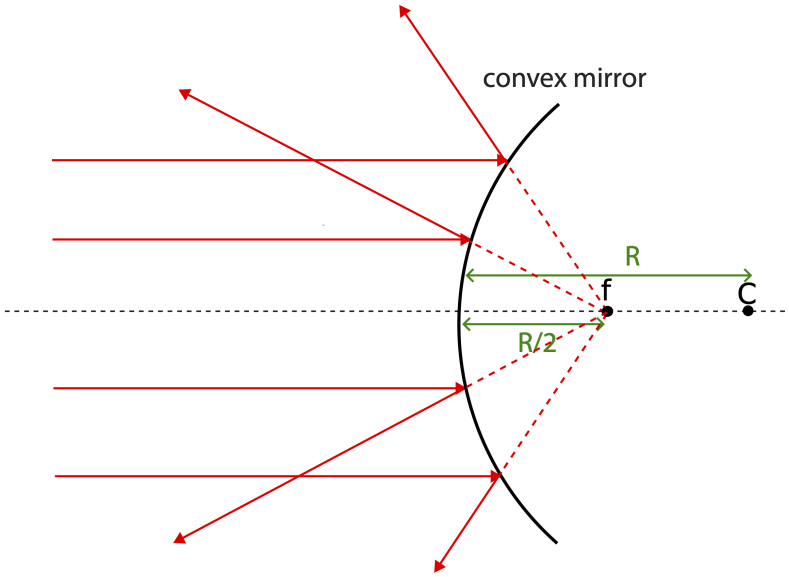
The big difference between the refected rays in this animation and those in Figure 10.3.3 is that the reflected rays detected by the observer no longer intersect. Instead, we find that they do cross each other behind the mirror when traced behind the mirror. The rays are not physically present behind the mirror, as indicated by the dashed lines, so the image formed is a *virtual image*, as described by the plane mirror at the start of this section. The image is also upright and enlarged, as opposed to inverted and reduced in size as in Figure 10.3.3.

Mathematically, the distinction between a real image and a virtual image made by a mirror is in the sign of the image distance. We will define a convention that distances in front of the mirror are always positive. Thus, the focal length and the object distance for the concave mirror are positive. This results in the image distance for an image formed in front of the mirror will always be positive, while the image distance for an image formed behind the mirror will be negative. This comes directly from Equation 10.3.6. Since both $f > 0$ and $o > 0$, the $1/i$ is negative when $1/f < 1/o$ or $o < f$ which is exactly the case in Figure 10.3.5 of an object placed between the focal point and the mirror. On the other hand, when $o > f$ Equation 10.3.6 guarantees that the image distance will be positive.

Convex Mirrors

What if the other side of the concave mirror was reflective, the side that curves away? This results in another type of spherical mirror, known as the *convex mirror*, as shown below. In this case incoming parallel rays from a distance object will *diverge* from each other when reflected, precisely because of the "curving out" shape of this mirror.

Figure 10.3.6: Convex Mirror



The convex mirror also has a radius of curvature, R, defined as the distance from the mirror to the center of the partial spherical shell. Except, in this case it is on the opposite side of the mirror due to its shape. We also notice that although the parallel incoming rays do not converge on the side of the mirror, when traced back, they meet at one specific point (as before, this is valid within the small angle approximation). We define this point as the focal point of a convex mirror. The focal length has the same relationship to the center as a concave mirror, $f = R/2$. However, since the the focal point does not focus physical rays, it is a virtual focal point. The same Equation 10.3.6 applies to convex mirrors.

As we did for concave mirrors, we can define three principle rays for a convex mirror. These are depicted in the animation below and described below:

- Principle ray #1: incoming ray parallel to the optical axis will reflect away from the focal point.
- Principle ray #2: incoming ray moving toward the focal point will reflect parallel to the optical axis.
- Principle ray #3: incoming ray moving toward center of curvature will reflect straight back.

Figure 10.3.7: Principal Rays of a Convex Spherical Mirror

As the animation shows the reflected rays will not converge in front of the mirror but will cross behind the mirror when traced back, forming a virtual upright image. In fact, a convex mirror will always make a virtual image of an object in front of the mirror. Using the sign convention defined when describing a concave mirror, the focal length will be negative for a convex mirror, since it is located behind it. From Equation 10.3.6 we can see that when the focal length is negative ($f < 0$) the image distance will always be negative ($i < 0$) since the object distance is always positive when an object is placed in front of the mirror ($o > 0$).

Below is the summary of the important sign conventions:

- distances measured to a location in front of the mirror are positive.
- distances measured to a location behind the mirror are negative.
- object distances are positive, $o > 0$.
- focal lengths of concave mirrors are positive, $f > 0$.
- focal lengths of a convex mirrors are negative, $f < 0$.
- image distances for real images are positive, $i > 0$.
- image distances for virtual images are negative, $i < 0$.

Magnification

We also want to mathematically analyze the size of images relative to the size of objects. The *magnification, M*, is defined as:

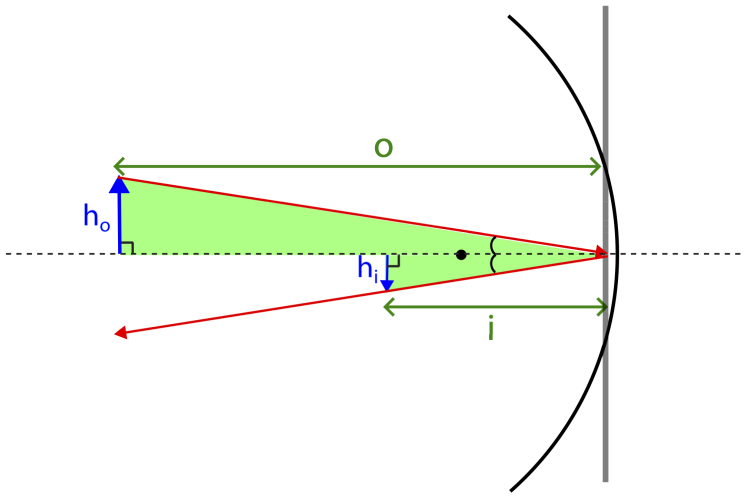
$$M = \frac{h_i}{h_o} \tag{10.3.7}$$

where h_i is the height of the image, and h_o is the height of the object. Similar to distinguishing between real and virtual images, we want to distinguish mathematically upright from inverted images. We do this by assigning a sign to the heights of objects and images. If we define the location of the optical axis as zero height, then any distance above the axis is positive, and a distance below the axis is negative. When a concave mirror creates a real image, it is inverted, as seen in Figure 10.3.3. The height of the image is negative since the inverted image is below the optical axis, resulting in negative magnification. On the other hand, when a mirror makes a virtual image which is upright, as seen both in Figure 10.3.5 and Figure 10.3.7, the magnification is positive since both the object and image are above the optical axis and have positive heights.

The magnitude of magnification tells us about the relative size of the image to the object. If the size does not changes, as for the plane mirror in Figure 10.3.1, then the magnification is one ($M = 1$), since the height of the object equals to the height of the image. If the image is larger than the object as in Figure 10.3.5, the object is *magnified*, and the absolute value of magnification is greater than one, since $|h_i| > h_o$. If the image is smaller than then object, as in Figure 10.3.3, the object is *demagnified*, and the absolute value of magnification (ignoring the image orientation) is less than one, since $|h_i| < h_o$.

Another interesting feature is the relationship of magnification to object and image distances, which can be demonstrated with a simple geometric argument. Consider the following triangles highlighted in Figure 10.3.8 below for a concave mirror creating a real image. To generate these triangle we used another "special ray" which is not one of the three principle rays. Since the optical axis (dashed horizontal line) goes through the center of curvature, it hits the mirror perpendicular to its surface. In other words, for the ray originating from the object below, the optical axis is the normal to the surface of the mirror. Thus, the reflected ray, that must goes to through the image, makes the same angle with the normal, according to the law of reflection. Therefore, since both triangles are right triangles and share another common angle, they are *similar* triangles. Thus, the ratio of the heights must equal to the ratio of horizontal distances marked.

Figure 10.3.8: Magnification Relationships



We just need to be careful about signs, since we defined the height of the image to be negative in this scenario. Thus, the ratio $h_i/h_o < 0$ by convention, but the ratio i/o is positive due to the sign conventions we established for object and image distances. In order to be consistent, we then define the relationship of magnification to object and image distances in the following way:

$$M = \frac{h_i}{h_o} = -\frac{i}{o} \tag{10.3.8}$$

where the minus sign in the equation assures the consistency of sign conventions. We can conclude by using the above equation, that for a positive object and image distance the magnification will be negative resulting in an inverted image, and for a positive object distance and a negative image distance, the magnification is positive resulting in an upright image.

Here is a summary of magnification sign and magnitude conventions:

- inverted images have negative magnetization, $M < 0$.
- upright images have positive magnetization, $M > 0$.
- magnified images have a magnitude of magnification greater than one, $|M| > 0$.
- demagnified images have a magnitude of magnification less than one, $|M| < 0$.

Example 10.3.2

A spherical shell is reflective on both sides. When the reflection of an object is viewed on the convex side, the image is 40% of the size of the object. If the shell is now turned around so that the reflection is viewed on the concave side, determine the size of the image (compared to the object), and whether the image is upright or inverted. Assume that the distance between the shell and object are unchanged after the shell is rotated.

Solution

Since we are working with one spherical shell where both side are reflective, the concave and convex sides will have the same radius of curvature and the same magnitudes of the focal lengths, except the concave side with have a positive focal length, while for the convex side the focal length will be negative. The object distance in this problem is the same for both the convex and concave sides. Since we know the magnification of the object when the convex side is used, we can relate the focal length to the object distance, then use this information to find the image distance in terms of object distance when the concave side is used.

Starting with the given information. The magnification when the convex side is used is:

$$M = -\frac{i}{o} = 0.4 = \frac{2}{5} \Rightarrow i = -\frac{2}{5}o$$

The magnification is positive since convex mirror can only make virtual upright images.

Solving for the focal length in term of the object distance:

$$\frac{1}{f} = \frac{1}{o} + \frac{1}{i} = \frac{1}{o} - \frac{5}{2o} = -\frac{3}{2o}$$

One check that there was no sign error is that the focal length for a convex mirror came out negative. The focal length when the concave side is used is then $f = \frac{2}{3}o$. Using this to find the image distance with the concave side facing the object:

$$\frac{1}{i} = \frac{1}{f} - \frac{1}{o} = \frac{3}{2o} - \frac{1}{o} = \frac{1}{2o} \Rightarrow i = 2o$$

Calculating the magnification:

$$M = -\frac{i}{o} = -\frac{2o}{o} = -2$$

Thus, the image is inverted and is double the size of the object.

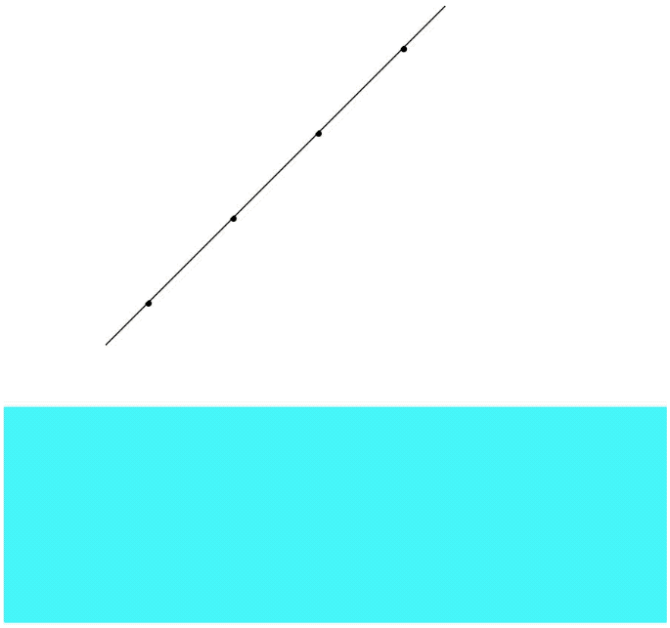
10.4: Refraction

Refraction

In the previous two sections we analyzed light rays changing direction through reflection, a process that occurs when a ray meets another medium. We now consider another way that direction change can occur, instead of reflecting from the medium, the ray moves into the new medium it encounters. This process, called *refraction*. To get to the essence of this phenomenon from Huygens's principle, we don't have a symmetry trick like we did for reflection, so rather than use a point source of the light, we can look at the effect that changing the medium has on a plane wave.

We saw in Figure 8.7.1 how a plane wave propagates according to Huygens's Principle. We can't sketch every one wavelets emerging from the infinite number of points on the wavefront, but we can sketch a few representative wavelets, and if those wavelets have propagated for equal periods of time, then a line tangent to all the wavelets will represent the next wavefront. It's clear that following this procedure for a plane wave will continue the plane wave in the same direction. But now let's imagine that such a plane wave approaches a new medium from an angle, as shown in the figure below. As each point on the wave front comes in contact with the new medium, it becomes a source for a new Huygens wavelet *within the medium*. These wavelets will travel at a different rate than they traveled in the previous medium (in the figure, the light wave is slowing down in the new medium). This means that the distance the wave in medium #1 travels is farther than it travels in medium #2 during the same time. The effect is a bending of the direction of the plane wave in medium #2 relative to medium #1.

Figure 10.4.1: Huygens's Principle Refracts a Plane Wave



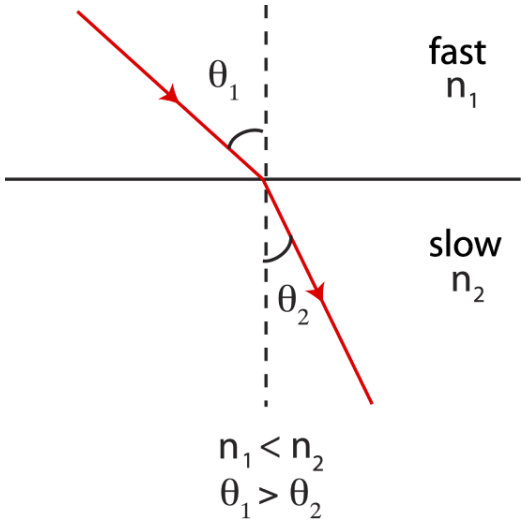
The amount that the direction of the light ray changes when the wave enters a new medium depends upon how much the wave slows down or speeds up upon changing media. The property of the medium which determines the speed of light is known as its *index of refraction*, n, and is defined as:

$$n = \frac{c}{v} \tag{10.4.1}$$

where c is the speed of light in a vacuum and v is a speed of light in the medium. When the light is traveling in a vacuum, $v = c$, the resulting index of refraction is one, $n = 1$. Since $v < c$ in any other medium except for the vacuum, the index of refraction will be greater than one, $n > 1$. In air light slows down by very slightly such that the index of refraction of air is, $n_{\text{air}} = 1.00029$. Since we will often analyze problems where light enters a new medium from air, for simplicity we assume that the index of refraction in air is approximately one. For other materials such as glass the index of refraction becomes significant, $n_{\text{glass}} = 1.52$.

Using ray representation refraction can be depicted as in the figure below, which depicts the same situation as in Figure 10.4.1, where light travels from a fast to a slow medium, or one with higher index of refraction to one with a lower one. As for reflection, we will measure angles relative to the normal of the surface separating two types of media. The incident ray makes an angle marked θ_1 with the normal, and the refracted ray makes an angle marked θ_2 with the normal. As demonstrated by Huygen's principle above, the rays bend down or toward the normal as they refract from a faster to a slower medium. Thus, in this example $\theta_2 < \theta_1$.

Figure 10.4.2: Ray Refraction Going from Fast to Slow Medium

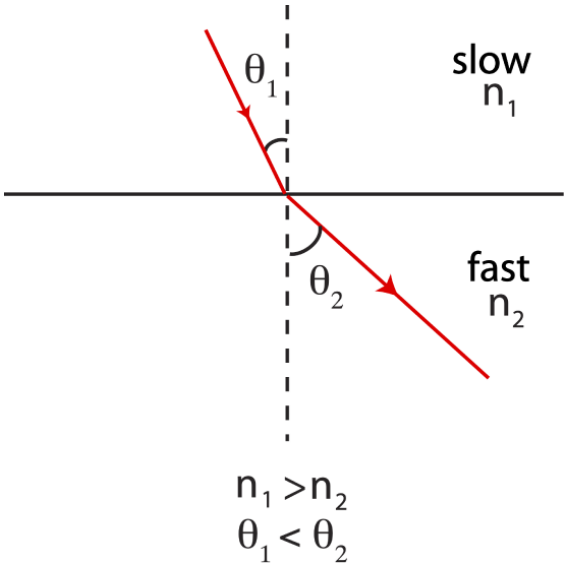


The relationship between the two angles and the two indices of refraction is described by *Snell's Law* (you can see the derivation of this equation in the digression below using Fermat's principle):

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{10.4.2}$$

Snell's Law is perfectly symmetric, which means that the same rules govern refraction when light moves from a slower to a faster medium. In other words, you can reverse the direction of rays shown in Figure 10.4.2, and you would obtain the situation depicted below. In this case the ray will bend away from the normal.

Figure 10.4.3: Ray Refraction Going from Slow to Fast Medium

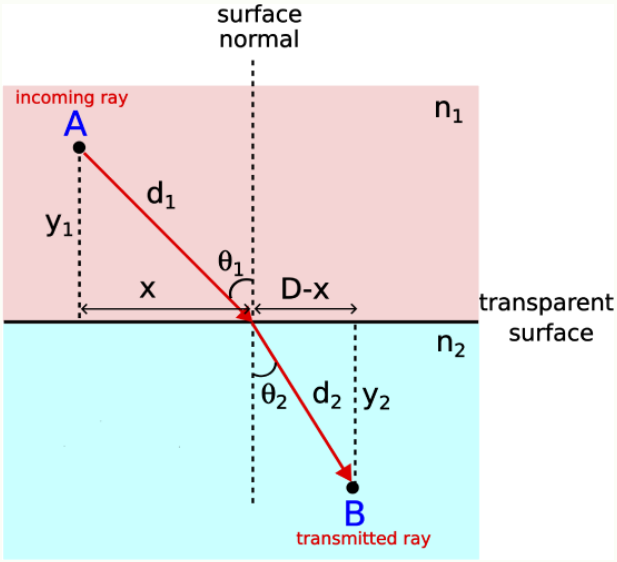


To summarize the general rules for refraction:

A ray will bend toward the normal when it moves from a faster to a slower medium, and it will bend away from the normal when it moves from a slower to a faster medium.

Digression: Snell's Law from Fermat's Principle

Fermat's principle states that light travels between two points such that it takes the shortest amount of time. Consider the diagram below. Here a incoming ray which originates at point A in a material with index of refraction n_1 is transmitted to another material with index of refraction n_2. We want to determine the path which will minimize the time of the ray's travel from point A to B.



Our goal is to find the the distance that the ray travels, $d_1 + d_2$, such that the time is minimized. As we did when we derived the Law of Reflection in Section 11.2, we will solve for the distance x, marked in the figure, such that it insures the shortest path between A and B with refraction.

The time it takes for the ray to travel from point A to point B is the time it takes to travel in medium 1 plus the time it travels in medium 2, $t_{\text{tot}} = t_1 + t_2$. The time in each medium is determined by the distance divided by speed, $v = c/n$:

$$t = \frac{d_1}{v_1} + \frac{d_2}{v_2} = d_1\left(\frac{n_1}{c}\right) + d_2\left(\frac{n_2}{c}\right)$$

Using Pythagorean theorem for each right triangle formed in the figure, the above equation can be written in terms of the horizontal distance x:

$$t = \sqrt{x^2 + y_1^2}\left(\frac{n_1}{c}\right) + \sqrt{(D-x)^2 + y_2^2}\left(\frac{n_2}{c}\right)$$

The vertical distances y_1 and y_2 and the horizontal distance D are fixed for given points A and B. To find the distance x which minimizes time, we need to take the derivative of time with the respect to x and set it to zero:

$$\frac{dt}{dx} = 0 = \left(\frac{x}{\sqrt{x^2 + y_1^2}}\right)\left(\frac{n_1}{c}\right) - \left(\frac{(D-x)}{\sqrt{(D-x)^2 + y_2^2}}\right)\left(\frac{n_2}{c}\right)$$

Expressing the above equation in terms of the angles θ_1 and θ_2 we get:

$$0 = \frac{n_1}{c}\sin\theta_1 - \frac{n_2}{c}\sin\theta_2$$

Rearranging, we arrive at Snell's Law:

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

Total Internal Reflection

It was noted above that light which passes from a slower medium to a faster one bends away from the perpendicular. What happens then if the incoming angle is made larger and larger (obviously it can't be more than 90^o)? For example, suppose we have $n_1 = 2.0$, $\theta_1 = 45^o$, and $n_2 = 1.0$. Plugging these values into Snell's law gives:

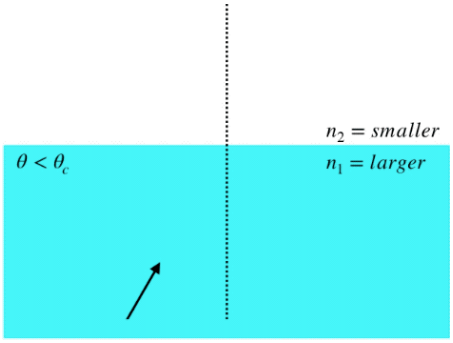
$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1 = 2.0 \cdot \sin 45^o = 1.4 \tag{10.4.3}$$

The sine function can never exceed 1, so there is no solution to this. This means that the *light incident at this angle cannot be transmitted into the new medium*. Every time light strikes a new medium some can be transmitted, and some reflected, so this result tells us that all of it must be reflected back into the medium in which it started. This phenomenon is called *total internal reflection*. The angle at which all of this first blows up is the one where the outgoing angle equals 90^o (the outgoing light refracts parallel to the surface between the two media). This angle is called the *critical angle*, and is computed by choosing the outgoing angle to be 90^o:

$$n_1 \sin\theta_c = n_2 \sin 90^o \quad \Rightarrow \quad \theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{10.4.4}$$

The process of increasing the incoming angle until total internal reflection is achieved is illustrated below.

Figure 10.4.4: Partial and Total Internal Reflections By Incident Angle



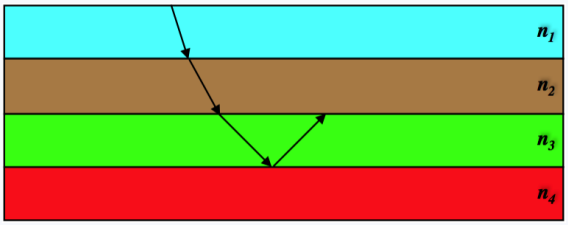
Note that there is at least partial reflection (obeying the law of reflection) every time the light hits the surface, but all of the light along that ray is only reflected when the ray's angle exceeds the critical angle.

Alert

Note that when light is coming from one medium to another, unless that light is a plane wave, it will be moving in many directions at once. Only the portions of the light wave with rays that equal or exceed the critical angle are not transmitted into the new medium. So the word "total" in "total internal reflection" to express the fraction of light at a specific angle that is reflected back, not necessarily the fraction of all the light that is reflected back.

Example 10.4.1

The diagram below shows the path of a ray of monochromatic light as it hits the surfaces between four different media (only the primary ray is considered – partial reflections are ignored). Order the four media according to the magnitudes of their indices of refraction.



Solution

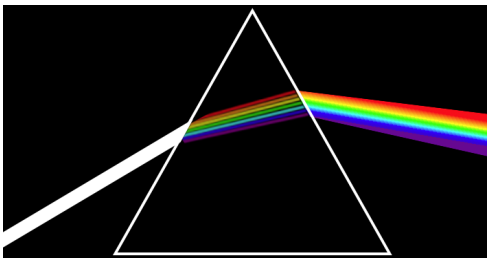
We know from Snell's Law that when light passes from a higher index to a lower one, it bends away from the perpendicular, so we immediately have $n_1 > n_2 > n_3$. For the ray to reflect back from the fourth medium, it has to be a total internal reflection (we are only considering primary rays, so this is not a partial reflection), which can only occur when light is going from a higher index of refraction to a lower one, so $n_3 > n_4$.

10.5: Dispersion

Dispersion

What determines the index of refraction for a medium is a very complicated problem in E&M, but there is one easily-observable fact: The amount that a ray bends as it enters a new medium is dependent upon the light's frequency. Specifically, the higher the frequency of the light, the more it bends – it essentially experiences a higher index of refraction when its frequency is higher. This phenomenon is most evident when white light is shone through a refracting object. The most iconic example of this is white light through a prism.

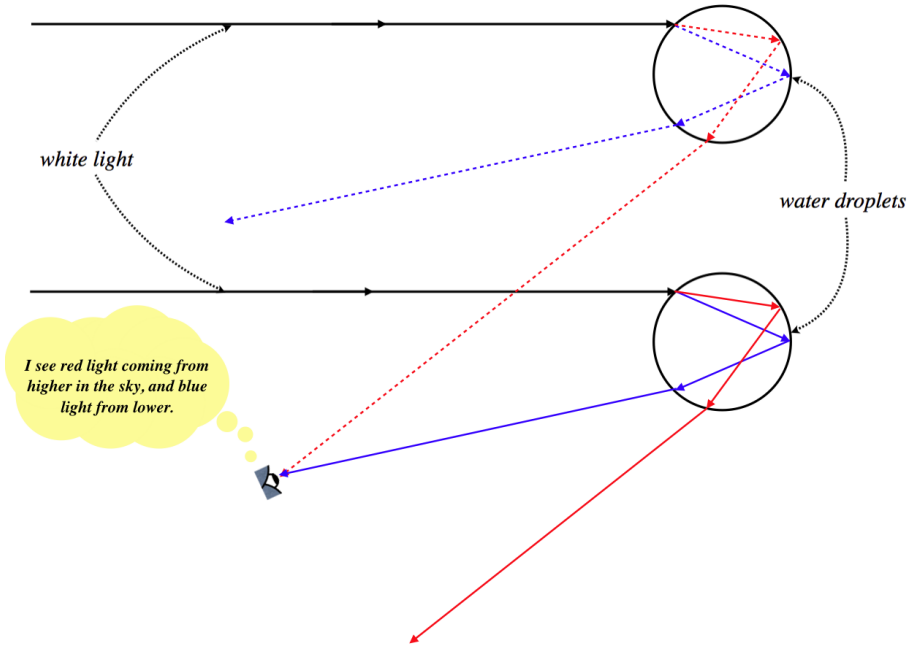
Figure 10.5.1: Dispersion Through a Prism



The emergence of the fully-separated spectrum of colors from a prism is reminiscent of a rainbow, and in fact rainbows are also a result of dispersion. Unlike the prism depicted above, however, internal reflection is an integral part of the rainbow effect (and in fact prisms can also feature internal reflection).

A droplet of water suspended in the atmosphere is a refracting sphere. White light that enters near the top of the droplet gets dispersed inside the droplet, reflects, and then gets dispersed as it exits the droplet, sending rays of different-colored light in different directions. The diagram below shows this effect for rays of red and blue light for two droplets.

Figure 10.5.2: Rainbows



A few things to note here:

- Notice that the sun always needs to be behind the observer in order to witness a rainbow. That's why it seems to move as you move, and why reaching the "end of the rainbow" is impossible (unless you can catch a leprechaun).
- The reason it is shaped like a bow is that the sun is nearly a point source, so the geometry is symmetric around the line joining the sun and the observer. If you create a "human-made rainbow" with a light and some mist, you can get close to an entire circle

(minus whatever light your body blocks out).

- The secondary rainbow above the primary one comes from the light that enters the *bottom* of the droplets, and has *two* internal reflections. This reversed direction of the light bouncing around inside the droplets results in the colors being reversed (the violet is at the top and the red at the bottom).

10.6: Lenses

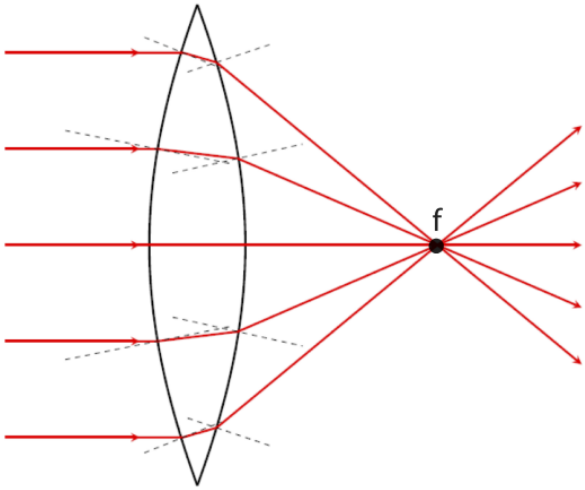
Lenses

In the previous couple of sections we saw that the law of reflection helps us understand how mirrors create images. In this section we will use the law of refraction to understand how another type of optical device, a *lens* can create an image. Instead of reflecting light like a mirror, a lens bends light, governed by the law of refraction, as it travels through a transparent material. There are numerous applications to lenses, the most common being corrective lenses uses in glasses to correct vision problems. In addition, lenses are used in technologies such as cameras, telescopes, and microscopes.

Converging Lenses

There are two types of lenses we will focus on. First type is a *converging lens*, which converges light upon refraction analogous to a concave mirror which converged light upon refection. The figure below show a particular shape of the converging lens, often called a *convex lens*, due to its shape.

Figure 10.6.1: Focal Point of Converging Lens

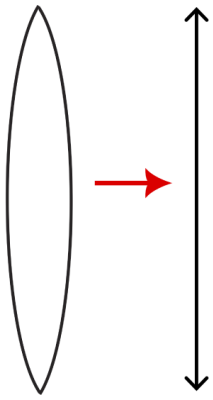


When parallel incoming light originating from a distance object or a laser refracts through the lens, the rays all converge to a single point, the focal point of the lens. You can see how this arises from refraction by following the bending of the rays twice. There are two refractions that occur, first rays travel from air to the glasslike material which bends light toward the normal. The second refraction occurs at the second boundary when the ray inside the lens refracts back into the air, this time bending away from the normal as it travels from a material of higher to lower index of refraction. This particular convex shape causes all the rays to converge to one point on the other side of the lens.

Unlike mirrors which have one reflecting side, lenses are perfectly symmetric. In other words, if you take the lens above and rotate it by 180º, the rays would converge to the same focal point. It's equivalent to saying that if you send rays from a distance object located on the right of the lens in the diagram above, the ray would converge to a point on the left, the same distance from the lens as the focal point on the right. Thus, unlike mirrors that have one unique focal point, lenses have two focal points on each side of the lens.

To simplify having to apply refraction twice, at each boundary we use the *thin-lens approximation*, which assume that the lens is infinitesimally thin compared to all other distances involved. Thus, we can assume that refraction only occurs once. Below is the symbol that we use to represent a converging lens in the thin-lens approximation.

Figure 10.6.2: Thin Lens Symbol for a Converging Lens



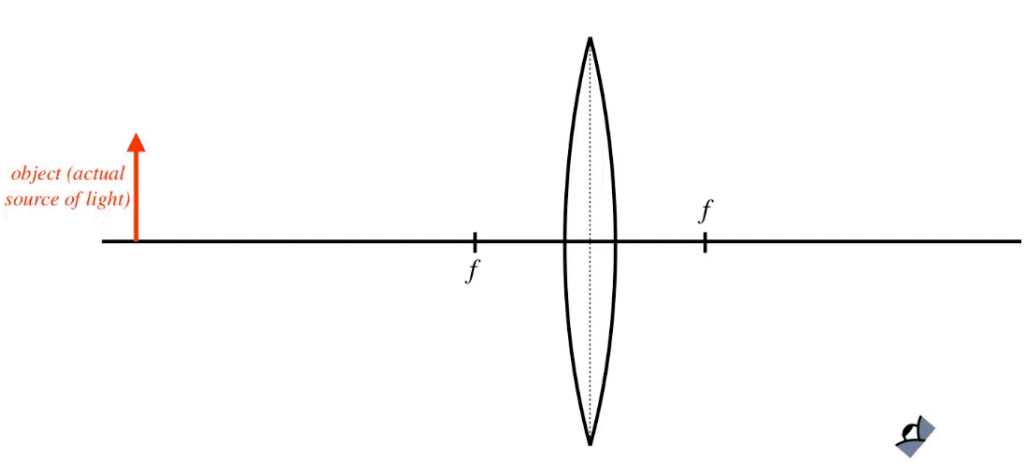
Now we would like to use ray tracing to determine the type of images a converging mirror makes. As for mirrors, we establish three principle rays for a converging lens. The principle rays are described and depicted in the animation below.

The three principle rays are:

- Principle ray #1: incoming ray parallel to the optical axis will refract through the far focal point.
- Principle ray #2: incoming ray that goes through the near focal point will refract parallel to the optical axis.
- Principle ray #3: incoming ray that goes through the center of curvature will follow a straight path as it goes through the lens.

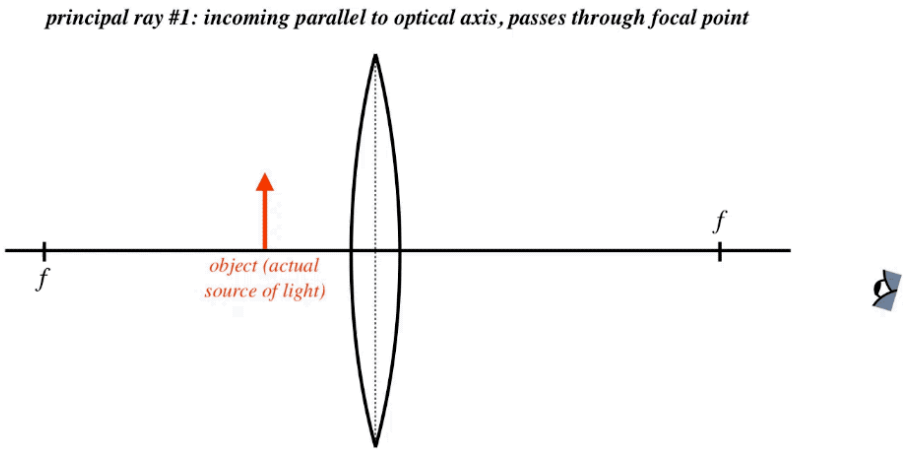
Although the animation below shows the entire shape of the lens, it is clear that only one refraction is demonstrated using the central vertical dashed line to represent the thin lens approximation. In this animation an object placed further from the lens than the focal point creates a real, inverted, and de-magnified image on the other side of the lens. The image is real since it is physical rays that go through the lens and converge on the other side.

Figure 10.6.3: Ray Tracing for a Converging Lens, object outside the focal point



These three principle rays change slightly when the object is placed closer to the lens than the focal point as shown in the animation below.

Figure 10.6.4: Ray Tracing for a Converging Lens, object inside the focal point



In this case, the rays do not converge as they refract through the lens. However, when traced back the appear to meet at a location in front of the lens (same side as the object), creating a virtual, upright, and magnified image.

We will not cover the derivations for lenses, but interestingly, when we use the thin-lens approximation, the same exact equations that emerged when using the small-angle approximation for mirrors applies to thin lenses:

$$\frac{1}{f}=\frac{1}{o}+\frac{1}{i};\quad M=\frac{h_i}{h_o}=-\frac{i}{o} \tag{10.6.1}$$

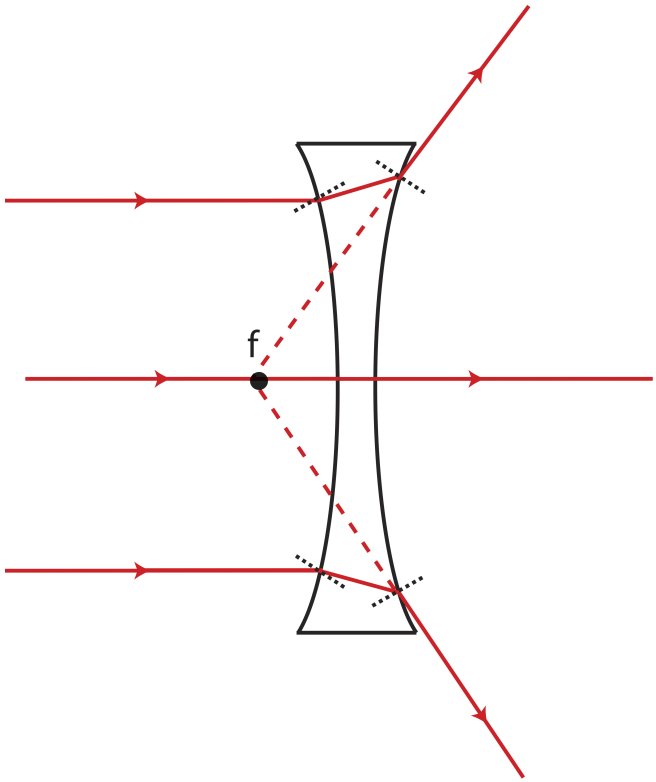
We will stick with the same convention as for mirrors: an optical device whose focal point is real will have a positive value for the focal length, such as the converging lens. Since object distances by convention are positive, this implies the image distance is positive for the scenario in Figure 10.6.3: $o>f$ implies $1/f>1/o$ resulting in $i>0$. When the object is placed closer than the focal point as in Figure 10.6.4 the image distance becomes negative: $o<f$ implies $1/f<1/o$ resulting in $i<0$. Another way of convincing yourself about the sign of the image distance is by looking at magnification. Since the real image in Figure 10.6.3 is inverted magnification has to be negative, resulting in positive image distance. On the contrary, for a virtual image, the magnification is positive, which means that the image distance in Equation 10.6.1 must be negative.

We conclude that the image distance is positive when it is real and formed on the other side on the lens. For concave mirrors, image distance was positive on the same side as the object when the image was real. Thus, the best way to remember the sign of the image distance is by thinking in terms of the image being real, implying $i>0$, or virtual which gives $i<0$.

Diverging Lenses

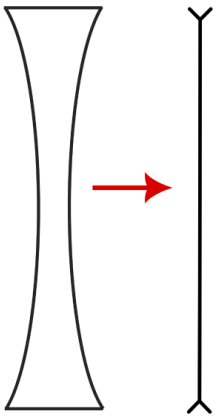
The analog to a convex mirror, is a *diverging lens*, also known as the *concave lens*, because of its shape shown below. When incoming parallel rays refract through a concave lens, they diverge away from each other on the other side of the lens. However, when traced back to the side of the lens from which the rays were incoming, we find they meet at a specific point. Thus, for diverging lenses the focal points are virtual, and will be negative, $f<0$, like the focal point for a convex mirror. As for converging lenses, diverging lenses have two focal points, one on each side of the lens.

Figure 10.6.5: Focal Point of Diverging Lens



The symbol for the thin-lens approximation of a diverging lens is shown in the figure below.

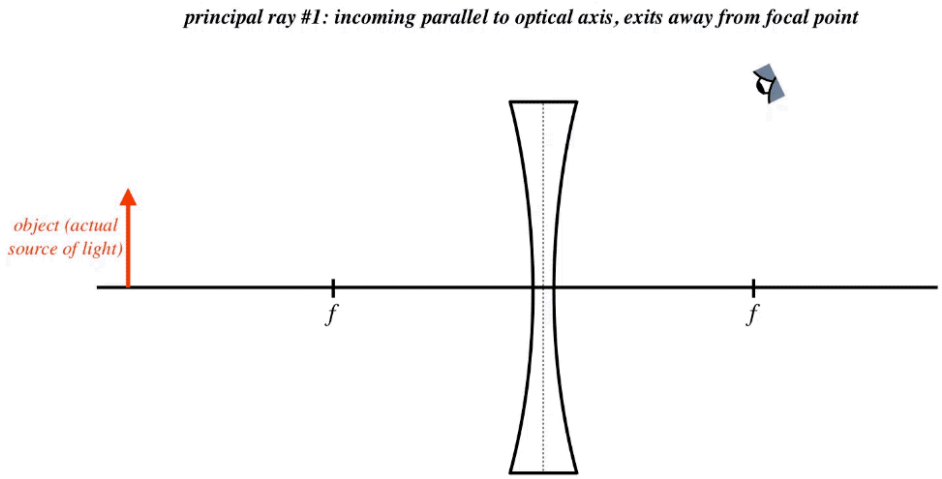
Figure 10.6.6: Thin Lens Symbol for a Diverging Lens



When assuming the thin-lens approximation we only consider refraction one time. The three principle rays for a diverging lens are described and depicted below:

- Principle ray #1: incoming ray parallel to the optical axis will refract away from the near focal point.
- Principle ray #2: incoming ray that goes toward the far focal point will refract parallel to the optical axis.
- Principle ray #3: incoming ray that goes through the center of curvature will follow a straight path as it goes through the lens.

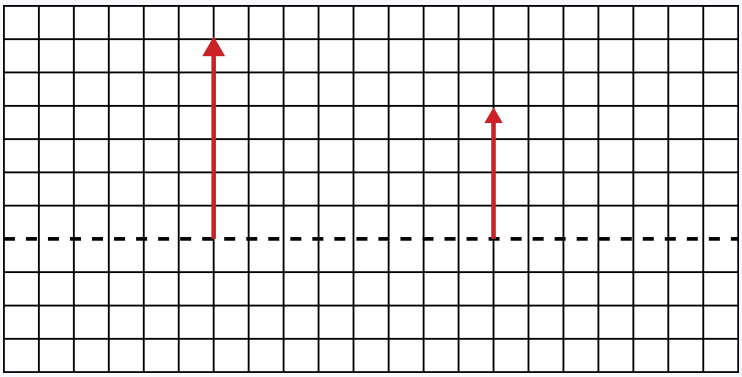
Figure 10.6.7: Ray Tracing for a Diverging Lens



The animation above shows a diverging lens making an upright, virtual, and de-magnified image in front of the lens. From Equation 10.6.1, we can see that diverging lenses will always form virtual images, since f is negative and o is positive. In addition, you can show using Equation 10.6.1 that the image distance will be always closer to the lens than the object, $|i| < o$. This result also implies that the image will always be de-magnified since $|i|/o < 1$.

Example 10.6.1

Below is a picture of two arrows. One of them is an object and one is an image. The heights and the distance between the two arrows are drawn to scale (each grid represents 1 cm in distance).



1. An optical device (a spherical mirror or a lens of either type) is placed in between the two arrows, resulting in the shorter arrow being the image.

   a) Is this a mirror or a lens and which type?

   b) Calculate the distance of the optical device from the object.

   c) Determine the focal length of the optical device.

2. You now use a converging lens. Determine which arrow is the image and which one is the object. Is the lens to the right, in between, or to the left of the two arrows?

Solution

1. a) Since both the object and the image are upright, the image must be virtual. Only a mirror produces virtual images on the other side of the object. Since the magnification is less than one, the mirror must be convex.

b) From the diagram, we can find the distance between the object and the image is:

$$o + |i| = 8cm$$

The magnification can also be measured from the diagram:

$$M = \frac{h_o}{h_i} = \frac{4}{6} = \frac{2}{3} = -\frac{i}{o}$$

This results in the following relationship between image and object distances:

$$|i| = \frac{2}{3}o$$

Plugging this back into the first equation:

$$o + \frac{2}{3}o = \frac{5}{3}o = 8cm \rightarrow o = 4.8cm$$

c) The image distance is negative:

$$i = -\frac{2}{3} \times 4.8cm = -3.2cm$$

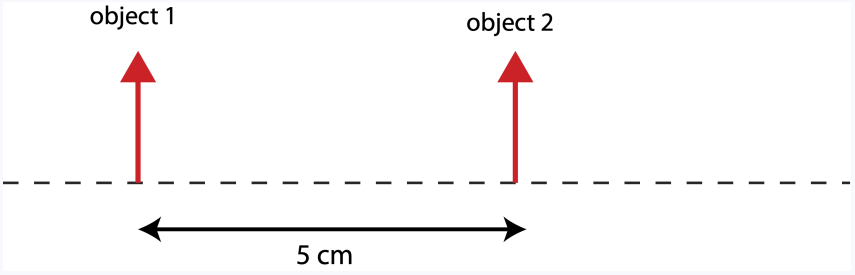
Plugging into the mirror equation:

$$f = \left(\frac{1}{4.8} - \frac{1}{3.2}\right)^{-1} = -9.6cm$$

2. When a converging lens produces a virtual image, the image is magnified. Thus, the taller arrow is the image and the shorter one is the object. Also, the object must be closer to the lens than the image, $o < |i|$, to give a negative image distance since the focal length is positive. Thus, the lens must be placed to the right of the two arrows, closer to the smaller arrow which is the object.

Example 10.6.2

Shown below are two identical objects (same size and orientation) which are 5 cm apart. A lens is placed 3 cm to the right of object 1. You observe that both images formed of the two objects are at the same location.



a) Determine the type of lens you have.

b) Calculate the location of the images in terms of distance to the right or left from object 1.

c) Determine the focal length of the lens.

d) Are the two images formed the same size? If not, which one is bigger?

Solution

a) For both images to be at the same location, one image must be virtual and the other one real. Thus, the lens is converging.

b) The image of object 1 is real since the lens is further from this object, so the object must be placed outside the focal length. The thin-lens equation for this case is:

$$\frac{1}{f} = \frac{1}{3} + \frac{1}{i}$$

The image of object 2 is virtual, the image distance is negative, but its magnitude is the same as for the image of object 1:

$$\frac{1}{f} = \frac{1}{2} - \frac{1}{|i|}$$

Setting the two equations equal and solving for i we get:

$$\frac{1}{3} + \frac{1}{i} = \frac{1}{2} - \frac{1}{i}$$

$$i = 12cm$$

The images are 15 cm to the right of object 1, 12 cm to the lens and another 3 cm to object 1.

c) Solving for focal length using the first equation in part b):

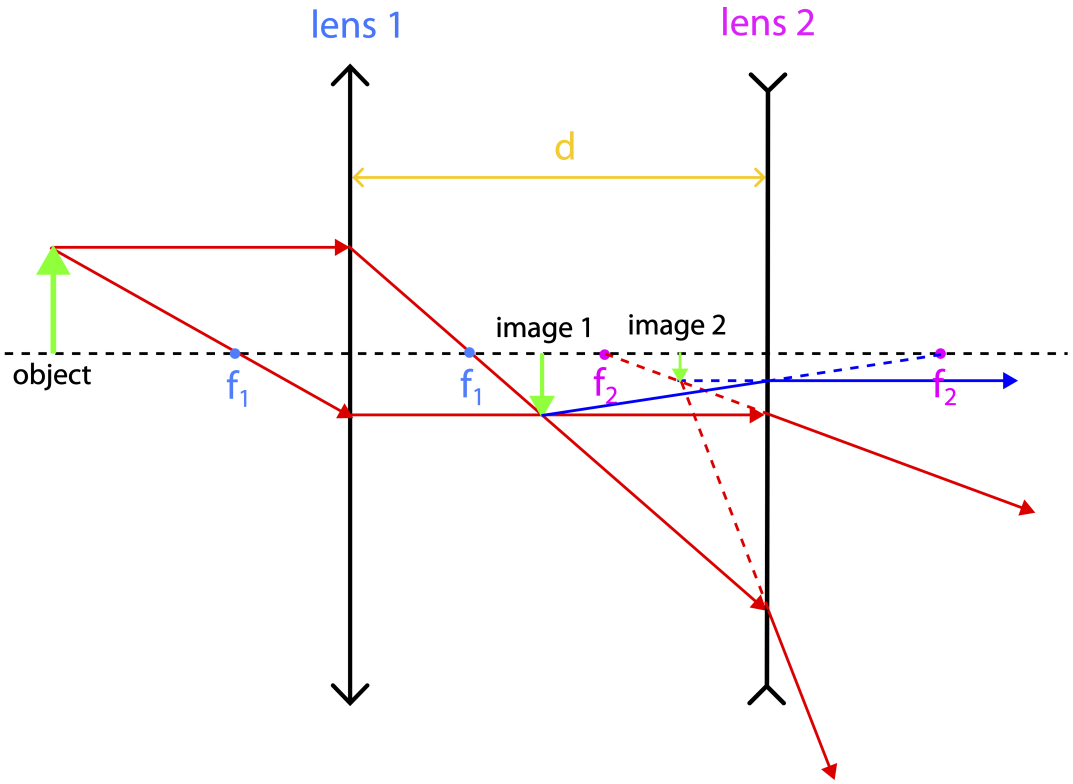
$$\frac{1}{f} = \frac{1}{3} + \frac{1}{12} = \frac{5}{12} \rightarrow f = 2.4cm$$

d) Since $M = -i/o$ and the image distance is the same for both, the one with the smaller object distance will be bigger, which is the image of object 2.

10.7: Multiple Optical Devices

Often to utilize the power of optical devices, multiple devices are used in combination to create images. The figure below shows an example of an object placed in front of two lenses, the first converging lens with focal point f_1 followed by a diverging lens with focal point f_2 placed a distance, d, apart. Our goal is to use tools of ray tracing, the thin-lens/mirror equation, and the magnification equation we applied to single optical device to determine image positions, types, and magnification of a multiple device system.

Figure 10.7.1: Two Lens System



In the figure light from the object will go through the first lens a governed by the single-lens ray tracing. In order to not overcrowd the image, only two of the three principle rays are shown above. The first lens is a converging lens, and since the object is placed further than the focal point, the image formed will be real, inverted, and on the opposite side of the the lens. This image is labeled "image 1" in the figure.

The rays that create the first image continue traveling and refract through the second lens. When red principle ray that refracts parallel to the optical axis encounters lens two, it is one of the principle rays for this lens as well, so we can readily determine how it will refract. Since lens 2 is a diverging lens, it will refract away from the near focal point. The other red principle ray from lens 1 is not one of the three principle rays for lens 2, so it is not easy to determine how it will refract.

However, there are many rays that pass through the first image and encounter the second lens. Some of these rays will be the principle rays for the second lens. Thus, we can draw the principle rays coming from image 1 to determine how lens 2 will form the final image. The blue ray is the second principle ray for lens 2, traveling toward the far focal point of lens 2, and refracting parallel to the optical axis. In other words, the *first image become the object for the second lens*. From these rays we can see that they will not converge behind lens 2, but can be traced back in front of lens 2 to create image 2.

The final image is an inverted virtual image. Although, one lens cannot make an inverted image which is virtual, this can occur with a multiple lens system. The first image was real and inverted, and the following virtual image keeps the orientation of the first image, so it remains inverted. For completeness, we show in the figure the other red ray that was not a principle ray for lens 2, refracting through lens 2. Since all rays participate in creating the image, the refracted ray will bend in such as way that it traces back to image 2, as shown.

Mathematically, we can treat this system as two independent single-lens systems. Initially, when the first lens makes the image, the location of that image can be calculated as follows:

$$\frac{1}{f_1} = \frac{1}{o_1} + \frac{1}{i_1} \tag{10.7.1}$$

where o_1 is the distance from the object to lens 1, f_1 is the focal length of lens 1, and i_1 is the distance of the image to lens 1. Next, we treat image 1 as the object for lens 2:

$$\frac{1}{f_2} = \frac{1}{o_2} + \frac{1}{i_2} \tag{10.7.2}$$

One thing that one must be careful about is connecting the image from lens 1 to the objects for lens 2. The image distance i_1 is the distance of the image from lens 1, while the object distance o_2 must be the distance of the object from lens 2. But the two quantities can be related when the distance d between the two lenses is known:

$$o_2 = d - i_1 \tag{10.7.3}$$

When i_2 is positive as in the case in Figure 10.7.1, then the object distance for lens 2 is simply the difference between the distance between the lenses and the distance of image 1 to lens 1. If the first lens made a virtual image on the same side as the original object of the lens, then the object distance to lens 2 would become the distance between the two lenses plus the distance of the first image to lens 1. Equation 10.7.3 would still apply since $i_1 < 0$ for virtual images, so the subtraction becomes an addition as expected.

For total magnification we want to compare the size of the final image to the size of original object:

$$M_{\text{tot}} = \frac{h_{i_2}}{h_{o_1}} \tag{10.7.4}$$

We would also like to determine how the total magnification relates to the object and image distances. To do this we first look at the magnification, M_1 by lens 1:

$$M_1 = \frac{h_{i_1}}{h_{o_1}} = -\frac{i_1}{o_1} \rightarrow h_{i_1} = -\frac{i_1}{o_1} h_{o_1} \tag{10.7.5}$$

Similarly the second magnification, M_2:

$$M_2 = \frac{h_{i_2}}{h_{o_2}} = -\frac{i_2}{o_2} \rightarrow h_{i_2} = -\frac{i_2}{o_2} h_{o_2} \tag{10.7.6}$$

The height of the first image become the height of the second object, $h_{i_1} = h_{o_2}$. Combing the two equations above we find:

$$h_{i_2} = \left(-\frac{i_2}{o_2}\right)\left(-\frac{i_1}{o_1}\right) h_{o_1} \tag{10.7.7}$$

And plugging back into Equation 10.7.4 for the total magnification we find:

$$M_{\text{tot}} = \left(-\frac{i_1}{o_1}\right)\left(-\frac{i_2}{o_2}\right) = M_1 \cdot M_2 \tag{10.7.8}$$

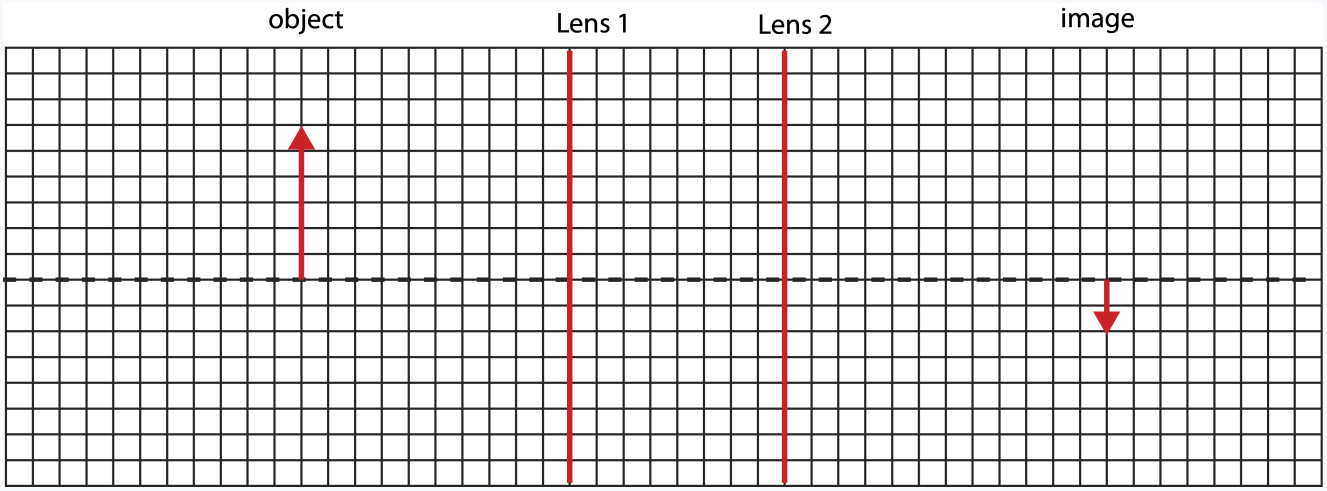
More generally, for a system of N-lenses the total magnification equation is:

$$M_{\text{tot}} = M_1 \cdot M_2 \cdot \ldots \cdot M_N = \prod_{n=1}^{N} M_n \tag{10.7.9}$$

If there are more than two optical devices, you follow the same procedure as you would for the two devices. The second image becomes the object for the third device, and so on. There are situations where an image by the first lens is made on the other side of the second lens. This is a situation of a virtual object which we will not cover in this course. Thus, we will only look at scenarios where the first image is created in front of the second lens.

Example 10.7.1

Bellow is a set up with two lenses of unknown types and focal lengths. The object and final image are shown. The heights of the object and image and all the distances are drawn to scale (each grid represents 1 cm in distance). Determine the type and the focal length of each lens.



Solution

The final image is real since it is on the other side of lens 2. Thus, lens 2 is converging. Since it's inverted from the original object, the first image had to be upright. Thus, the first image is virtual, which means that lens 1 can be either diverging or converging. So we need to do further calculations to determine the nature what type is lens 1.

Let's first gather what information is given in the figure. The object is 10 cm away from lens 1: $o_1 = 10cm$. The two lenses are 8 cm apart: $d = 8cm$. The final image is 12 cm to the right of lens 2: $i_2 = 12cm$. Also, the height of the object is 6m, and the height of the final image is -2 cm. From the heights we can determine the total magnification is:

$$M_{tot} = \frac{h_{i_2}}{h_{o_1}} = -\frac{2}{6} = -\frac{1}{3} = \left(-\frac{i_1}{o_1}\right)\left(-\frac{i_2}{o_2}\right) = \left(\frac{i_1}{10cm}\right)\left(\frac{12cm}{8cm - i_1}\right) = \frac{6}{5}\cdot\frac{i_1}{8 - i_1}$$

From the above equation we can solve for i_1:

$$5(i_1 - 8) = 18i_1 \rightarrow i_1 = -\frac{40}{13}cm$$

Since we know the object distance for lens 1, we can now solve for the focal point of lens 1:

$$\frac{1}{f_1} = \frac{1}{o_1} + \frac{1}{i_1} = \frac{1}{10} - \frac{13}{40} \rightarrow f_1 = -\frac{9}{40}cm$$

Thus, lens 1 is diverging since the focal length is negative. Since we know the location of image 1, we can find the location of object 2:

$$o_2 = 8cm + \frac{40}{13}cm = \frac{144}{13}cm$$

Finally, solving for the focal length of lens 2:

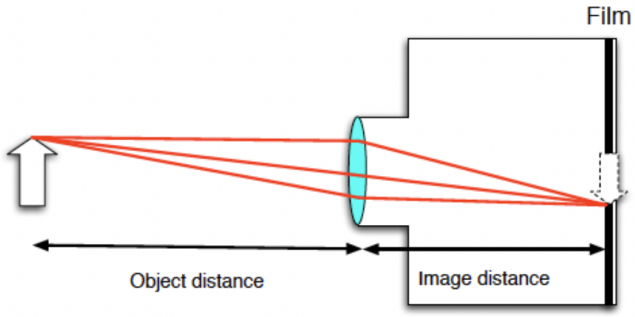
$$\frac{1}{f_2} = \frac{1}{o_2} + \frac{1}{i_2} = \frac{13}{144} + \frac{1}{12} \rightarrow f_2 = 5.76cm$$

10.8: Applications

Cameras

A camera which takes pictures can be explained using basic principles of geometric optics. The goal is to capture light coming from a physical object and project a de-magnified image of the object on a screen inside the camera. This can only be achieved by a converging lens that can create a real image on the screen. Here we model a camera as a hollow box with film or a detector on the back wall, which acts as our screen, where the image is detected. 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 digital) work.

Figure 10.8.1: Toy Model of a Camera



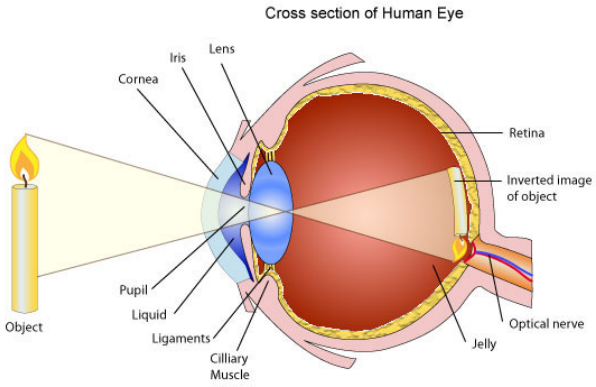
Note that since the camera lens produces a real image, it will appear upside-down on the film negative. This is taken into account in the film developing and printing process; this negative is used to project yet another real image onto the photograph print. This print is then right-side-up, unless your 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, so that the image is in focus. In fact, if you inspect the thin lens equation, as the object distance o decreases (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.

The Eye

The human eye processes images in a very similar manner as a camera. Light from an object refracts through the lens of our eye, and a real inverted image gets projected unto the retina in the back of the eye. Then the information travels through the optical nerve to our brain which interprets this image. A basic illustration of the anatomy of the eye is shown below. In contrast to a camera we cannot change the image distance significantly for our eye, since the length of the eye is fixed.

Figure 10.8.2: Anatomy of an Eye



To get a focused image, the image distance i must be the same as the diameter of our eyeball. For most people this distance is roughly 2.5 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 in focus at a variety of distances? Unlike camera lenses, the lenses in our eyes can change focal length. 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 lenses' shape, thus changing its focal length.

When the ciliary muscles are relaxed the lens is (relatively) flat, and the light rays are not bent much as they pass through. This results in a larger focal length, because it would take a long distance for light rays parallel to the optical axis to converge to a point. We relax our eyes when looking at distant objects, when our lens is at its *longest focal length*, f_{max}. When the ciliary muscles contract, the lens becomes more round, and the normals change more. This corresponds to more bending of the light as it passes through the lens; the rounder lens has a smaller focal length. This minimal focal length is used when looking at near objects. Our ability to change the focal length of our eyes is referred to as *accommodation*.

There is a limit to how much the lenses in your eyes can change shape. Consequently there is a *shortest focal length*, f_{min}, that your eyes can have, and a closest object that you can focus on clearly. The nearest distance that you can hold an object while still clearly focussing on it is called your *near point*, d_{np}. Note that although it is called a "point" we actually refer to a distance. It is not the same as the shortest focal length, f_{min} rather they are related by:

$$\frac{1}{f_{min}} = \frac{1}{d_{np}} + \frac{1}{i} \tag{10.8.1}$$

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 the near point of a middle-aged person is around 25 cm.

Similarly there is a furthest distance you can focus on when you totally relax your eyes. This distance is known as your *far point*, d_{fp}. For “normal” healthy eyesight the far point is infinity – there is no furthest distance someone can focus on. However, if your eyes cannot relax completely, or your relaxed focal length is longer than your eyeball’s diameter, then you will have a far point. That is, you will not be able to focus on objects beyond some distance d_{fp}. If your far point is at infinity, this means that the maximum focal length of a healthy eye is around the image distance, $f_{max} \sim 2.5cm$.

Corrective Lenses

Let us 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 10 cm, out to 25 cm 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 25 cm, glasses or contacts are prescribed to correct this vision defect.

Farsightedness (or *hyperopia*) is a condition where only far objects can be seen clearly. This is either because the lens cannot become round enough, or because the distance between the retina and the lens is too short. A farsighted person can see distant objects just fine with slight accomodation. As objects get closer, the eye must accommodate more strongly 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 25 cm, then glasses or contacts are prescribed to correct this vision defect.

Nearsightedness (or *myopia*) is the condition where only nearby things can be seen clearly. 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 "infinity", 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.

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 responds the same to light that comes off an object directly and light that appears to be coming from the image of some object (as would be the case when we are wearing corrective lenses).

This gives us a strategy for modeling 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.

We can approximately model the required prescription (focal length) of the corrective lenses using the concepts introduced here. To correct presbyopia or hyperopia the corrective lens needs to make an image of the object at the near point of the patient's eye, when the object is at the near point of a healthy eye, $o = 25\text{cm}$. For example, when reading you would like to hold the book at 25 cm, and your corrective lenses will make an image of the book at your near point where your eye can focus. The image has to be virtual since it needs to be created in front of the corrective lens. Also, since the image distance is further than the object distance, the lens must be converging. Neglecting the distance between the glasses and the eye, the focal length of the glasses is then:

$$\frac{1}{f_{lens}} = \frac{1}{0.25\text{m}} - \frac{1}{d_{np}} \tag{10.8.2}$$

The object distance is written in meters due to units used for lenses in optometry, as described below. Assuming that near point is a positive distance, the minus sign in the equation above assures that the image distance is negative, since the image created is virtual.

For myopic eyes, the corrective lens needs to make an image of an object at "infinity", $o = \infty$, at the far point of the patient where their eyes can focus. This lens must be diverging since the virtual image is closer to the lens than the object. Since $1/\infty = 0$, the focal length of the corrective lens (again neglecting the distance between the lens and the eye) is:

$$\frac{1}{f_{lens}} = -\frac{1}{d_{fp}} \tag{10.8.3}$$

As for the case of converging corrective lenses, the minus sign in the equation above assures that the image distance is negative since the image is virtual and the far point is a positive distance. This directly results in a negative focal length for the diverging lens being used to correct myopia.

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 is the inverse of the focal length, 1/f. The SI units of *diopters* (D) which is 1/meter. Thus, when converging a lens's focal length to diopters make sure the focal length is in units of meters.

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 converging ($f > 0$) or diverging ($f < 0$) lenses are needed, the prescriptions for the lenses can be either a positive or negative number of diopters.

Example 10.8.1

On a planet in a galaxy far, far away aliens have eyes with a “normal” length of 3.5cm. An alien named Asil needs reading glasses, because her eyes are shorter than the "normal" length. She was prescribed reading glasses with an optical power of 2.75D to correct her vision (so she can focus on objects as close as the aliens with “normal” eyes can). The lenses in Asil's eyes have the same minimum focal length, of 2.8cm, as that of "normal" eyes on her planet.

a) What is the closest distance a person with “normal vision” (on her planet) can focus on?

b) Find the length of Asil’s eyes.

Solution

a) The minimum focal length, $f_{\min}$, of the eye is when an object is placed at the near point, $o = N_{\text{normal}}$, and a real image is formed at a distance of the length of the eye, $i = L_{\text{normal}}$:

$$\frac{1}{f_{\min}} = \frac{1}{N_{\text{normal}}} + \frac{1}{L_{\text{normal}}}$$

Solving for the near point of a healthy eye:

$$N_{\text{normal}} = \frac{1}{\frac{1}{2.8cm} - \frac{1}{3.5cm}} = 14cm$$

b) We can find Asil’s eye length by thinking that that when an object is at her near point, $o = N_{\text{Asi}}$, the image is at the length of her eye, $i = L_{\text{Asil}}$, when her eye’s focal length is at its minimum:

$$\frac{1}{f_{\min}} = \frac{1}{N_{\text{Asil}}} + \frac{1}{L_{\text{Asil}}}$$

Solving for the length of the eye:

$$L_{\text{Asil}} = \frac{1}{\frac{1}{f_{\min}} - \frac{1}{N_{\text{Asil}}}}$$

To find Asil’s near point we use the equation for reading glass prescription: place object at the desired near point of 14cm (found in a) to create a virtual image at Asil’s near point: $i = -N_{\text{Asil}}$:

$$\text{Optical Power} = 2.25D = \frac{1}{N_{\text{normal}}} - \frac{1}{N_{\text{Asil}}}$$

Solving for Asil's near point (we need to convert diopters to centimeters $D = 1/meter$:

$$N_{\text{Asil}} = \frac{1}{\frac{1}{14cm} - \frac{0.0275}{cm}} = 22.8cm$$

Finally, solving for the Asil's eye length using the first equation above:

$$L_{\text{Asil}} = \frac{1}{\frac{1}{2.8cm} - \frac{1}{22.8cm}} = 3.19cm$$

Asil's eye is shorter than a healthy eye length on this planet. This implies that light focuses behind Asil's eye and not on her retina, so she needs a converging corrective lens that would focus the light at a shorter distance, or on her retina. This is exactly why she needs reading glasses which contain converging lenses.

10.9: Summary

Overview of Mirrors and Lenses

Table 10.9.1: Summary of properties of mirrors and lenses

focal point and device description	focal length sign	image location and type	image distance sign	orientation	magnification
Concave mirror or converging lens	f>0	if o>f: a real image front of the mirror or on the other side of the lens compared to the object if o<f: a virtual images behind the mirror or on the same side of the lens as the object	if o>f, then i>0 if o<f, then i<0	if o>f, the image is inverted, M<0 if o<f, the image is upright, M>0	if f<o<2f, the image is magnified, |M|>1 if o>2f, the image is de-magnified, |M|<1 if o<f, the image is magnified, |M|>1

lens is on the same side of the lens as the object and both sides of the lens				
f<0 for a diverging lens	virtual image behind the mirror or on the same side of the lens as the object for all object distances	i<0 for all o's	the image is upright, M>0 for all o's	all images are de-magnified, |M|<1

g l h n s m i r r o r a n d b o t h s i d e s o f t h e l e n s				

Ray Tracing

Concave mirror: object further(closer) than focal point:

- Principle ray #1: incoming ray parallel to the optical axis will reflect through the focal point.
- Principle ray #2: incoming ray that goes through (away from) the focal point will reflect parallel to the optical axis.
- Principle ray #3: incoming rays that goes through (away from) the center of curvature will reflect straight back.

Conex Mirror:

- Principle ray #1: incoming ray parallel to the optical axis will reflect away from the focal point.
- Principle ray #2: incoming ray moving toward the focal point will reflect parallel to the optical axis.
- Principle ray #3: incoming ray moving toward center of curvature will reflect straight back.

Converging Lens: object further(closer) than focal point:

- Principle ray #1: incoming ray parallel to the optical axis will refract through the far focal point.
- Principle ray #2: incoming ray that goes through (away from) the near focal point will refract parallel to the optical axis.
- Principle ray #3: incoming ray that goes through the center of curvature will follow a straight path as it goes through the lens.

Diverging Lens:

- Principle ray #1: incoming ray parallel to the optical axis will refract away from the near focal point.

- Principle ray #2: incoming ray that goes toward the far focal point will refract parallel to the optical axis.
- Principle ray #3: incoming ray that goes through the center of curvature will follow a straight path as it goes through the lens.

CHAPTER OVERVIEW

11: Electromagnetism

11.1: Fields

Introduction

A *field* is a concept that is used in almost every part of physics. In Physics 7C we will concentrate on the *electric* and *magnetic* fields. The concept of fields will help us understand how electric charges interact with each other and how those interactions lead to electric and magnetic forces. In the last section we will see electric and magnetic fields are closely related and can propagate: a phenomenon we commonly refer to as "light"! Thus, we will see that fields are not simply concepts that help us explain these interactions, but fields are indispensable: neglecting them would lead to a violation of both the conservation of energy and conservation of momentum.

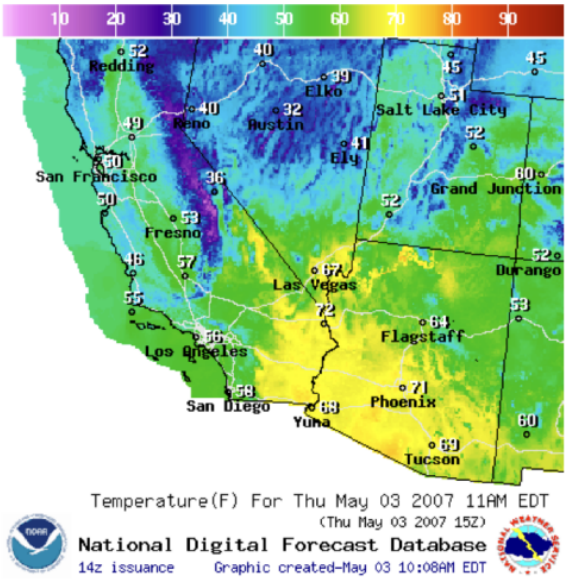
We will first start by discussing the potential energy that results from charges interacting, and see how that connects to electric forces. From this will will construct the *potential field* and the *electric field*, and develop the relationship between these fields and potential energy and the electric force. We will follow with the introduction of the *magnetic field* and *magnetic force* as we connect magnetism to electricity.

The idea of a field is rooted in the concept that there is some physical quantity that has a value “everywhere". The value 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 vary in time as well as space, 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 of fields.

Scalar Fields

A *scalar field* is a field which describes a scalar quantity in space and time. The weather map below is similar to one you might see on the news: it represents a field. There is not one universal temperature, so you need to define both a time and a position where that temperature can be found. A question like “what is the temperature in San Francisco now?” can be answered because we have specified both the place and the time. In other words, the temperature T is a function of position (x, y, z) (for example), and time t.

Figure 11.1.1: Weather Map as a Scalar Field Representation



Courtesy of the National Weather Service

The weather map above is similar to one you might see on the news; it represents a field. There is not one universal temperature; the if you want to reference a temperature, you need to define both a time and a position where that temperature can be found. A question like “what is the temperature in San Francisco now?” can be answered because we have specified both the place and the time. In other words, the temperature T is a function of position (x, y, z) (for example), and time t.

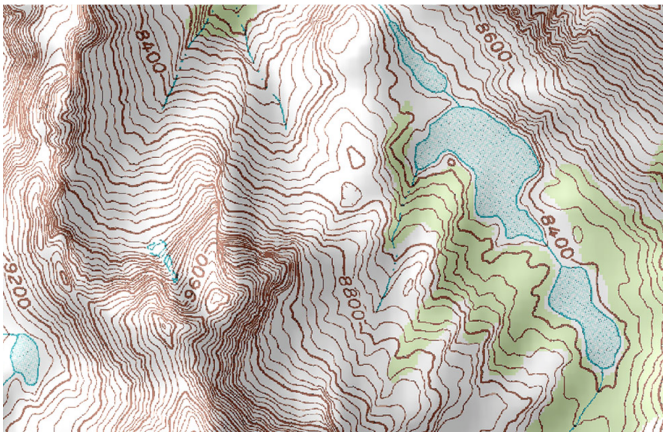
With that being said, let's note a few important things about the map above:

- It shows temperatures at a particular time, so what is being shown is $T(x, y, t = \text{today})$. The full field $T(x, y, t)$ could be represented by an entire archive of all previous (and future!) temperature maps.
- On this map, the temperature is represented as a color, with the scale above giving the corresponding values.
- On some weather maps the temperature is only shown for selected locations. Even in the places where a temperature is not labeled, a temperature can be measured. There is a defined T for every shown value of (x, y).

Topological Fields

A *topography map* is another example of a scalar field representation. The map shown below represents the height of the Earth's surface as a function of position. Because the Earth does not shift quickly, we can neglect that this map depends on time. This map mainly displays height by drawing lines along paths of equal height. The land along a single contour is at the same elevation. Neighboring lines are separated by the same difference in height, which gives us information about the steepness of an area. When the density of lines is large (lines are close together), this means that the region is very steep since the height is changing quickly with short distances. This typically represents a peak of a mountain as in the region of about 9600 feet in the map below. Regions of widely separated lines represent flat areas such as valleys as seen near the body of water below. Drawing a line in a scalar field such that every point on the line has the same value in the field is an idea that will be utilized more later.

Figure 11.1.2: Topological Map as a Scalar Field Representation

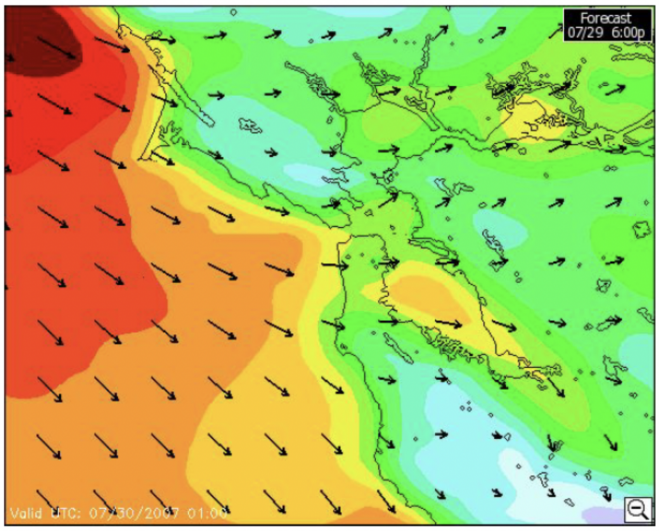


Vector Fields

The previous examples were scalar fields because they describe scalar quantities (like temperature or height). *Vector fields* define a vector, that has both a *magnitude* and *direction*, for all positions and times. Vector fields are much more relevant to the topics discussed later in this chapter, so take a while to make sure you can distinguish between the two.

A wind map shown below displays the velocity of the wind at various locations at a fixed time. Because the velocity is a vector, this map must display both the direction and magnitude of wind. The color is used to represent areas of the map where the wind is a fixed magnitude. The arrows show the magnitude and direction of wind at various locations. Like before, although the arrows aren't drawn everywhere on this map, the field defines a magnitude and direction for *every* point in the map.

Figure 11.1.3: Wind Map as a Vector Field Representation



Map courtesy of WeatherFlow, Inc.

The representation of a vector field presented above, where arrows display the vector field at various locations, is called a *field map* representation. We will become familiar with other ways to represent vector fields, each of which with its own advantages and disadvantages. The advantage of field maps is that they can be read by eye very quickly, but it omits information about the vector field's scale.

Alert

There may be examples of functions that fit into our definition of "a field" which are probably not very useful. For example, one could define an elephant field in the following way: given a specific position and a specific time, the elephant field describes how many elephants exist there. Certainly this "field" fits into our definition, but it is not a terribly *useful* thing to consider. We cannot use the "elephant field" to make any new predictions, it is not required by any experiments, and it does not simplify any discussions. It is important to recall that we only consider something a field where it is required by experimental evidence (like the electric and magnetic fields) or it is convenient for discussing the quantity (like temperature).

11.2: Electric Force

Electric Charge

Electric charges are measured in *Coulombs*, abbreviated *C*. The fundamental charge, e is given by:

$$e \equiv 1.602 \times 10^{-19}\ \text{C} \tag{11.2.1}$$

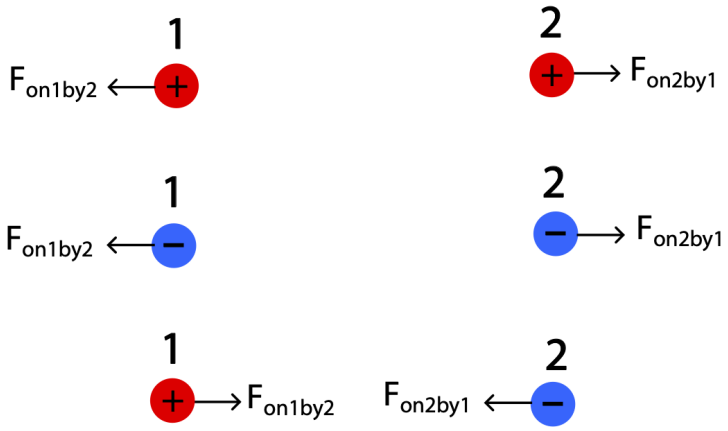
The electron, e^-, has a charge of $q = -e = -1.6 \times 10^{-19}$ C . The proton, p^+, has a charge with the same magnitude but it's positive, $q = +e = 1.6 \times 10^{-19}$ C . All atoms, molecules, or charged macroscopic objects get their charge from either an excess or deficit of electrons compared to the number of protons they have: a charge of "e" is thus considered fundamental, as all observed particles or objects with charge have some integer multiple of this value. For example, if a macroscopic object has an excess of 6.24×10^{19} electrons, then that object has a charge of -10 C.

Electric charge is conserved, it cannot be created out of thin air or disappear forever. For example, when you charge an object by rubbing it with a cloth, then the amount of charge that you add (or remove) to the object is the amount of charge that gets removed (added) from the cloth. The way charges get transferred between objects highly depends on the physical properties of those objects. Two distinct categories are *conductors*, objects where charges are free to move, and *insulators*, objects where charge is locked in place.

Electric Force

An *electric force* is an interaction between two electric charges. It is one of the fundamental forces in nature, which we will later combine with magnetism to describe the *electromagnetic force*. The force depends on the sign of the charges, the magnitude of the charges, and the distance between them. The figure below shows forces for the three possible pair combinations of charges, emphasizing that the electric force, like all other forces, is a pairwise interaction.

Figure 11.2.1: Forces Between Charges



The forces are governed by Newton's third law, such that the force of charge 1 by charge 2 is equal in magnitude but opposite in direction to the force on charge 2 by charge 1. We see from the figure that when both charges are the same, the force wants to push them apart. When the two charges have opposite sign, the electric force wants to bring them together. This is a fundamental property of the electric force: *like charges repel and opposite charges attract*.

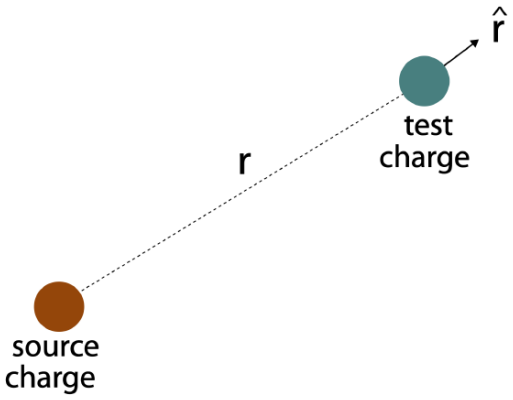
The magnitude of the electric force between two charges q_1 and q_2, with their centers separated by a distance r, is given by the following equation:

$$F_{\text{on}q_1\text{by}q_2} = \left|\frac{kq_1q_2}{r^2}\right|;\ k = 9 \times 10^9 \frac{\text{Nm}^2}{\text{C}^2} \tag{11.2.2}$$

The constant k converts force to the proper units of Newtons. The force increases linearly with the magnitude of each charge, but decreases as the inverse of the distance squared. Thus, the force become weak quickly as the charges separate. Newton's third law guarantees that the force exerted on q_2 by q_1 has the same magnitude as the force above but will point in the opposite direction.

The absolute value is placed there to insure that we are calculating the magnitude of the force regardless of the signs of the charges. However, force is a vector, so we also need to consider the direction of the force. Let is distinguish the two interacting charges, as one being the charge that generates the force, the *source charge*. The other charge that feels the force is the *test charge*. We then define a vector that points radially (along the distance separating the two charges) away form the source charge at the location of the test charge. The vector that specifies the direction is $\hat{r}$, where the "hat" above the "r", means it is a *unit vector*, which has no magnitude (and thus no units). It is a vector that specifically specifies the direction away from the source charge.

Figure 11.2.1: Direction of Electric Force



We then write down the force vector in the following way, known as *Coulomb's Law*, which describes an interaction between point charges:

$$\vec{F}_{\text{on} q_1 \text{by} q_2} = \frac{k q_1 q_2}{r^2} \hat{r} \tag{11.2.3}$$

Let us now check that this equation is consistent with properties of interacting charges. From the above equation you can see that whether both q_1 and q_2 are both positive or both negative, the quantity $q_1 q_2$ will be positive, so the force vector points along the radial unit vector. Thus, the force points away from the source charge. This force is then *repulsive*, since the two charges repel each other. When the two charges have opposite signs, then the quantity $q_1 q_2$ is a negative, so the force vector points in the opposite direction of the radial unit vector (toward the source charge). Therefore, the two charges are drawn toward each other, and the force is *attractive*.

11.3: Electric Field

Electric Field

In the previous section we made an effort to distinguish the source charge that "generates" the force, and a test charge that "feels" the force. Imagine the source charge is fixed in space, and you move the test charge to another location near the source charge. Then it would will a different force in the new location. In fact, anywhere in space around the source charge, the test charge feels a force. Thus, we can think of the source charge generating a vector quantity everywhere in space, which we will call the *electric field*. This is consistent with our definition of a vector field from Section 11.1, a vector quantity which has a value everywhere in space. Then, a test charges feel as a force when brought into this field. In other words the source charge generates a vector field around it which can be detected by test charges in a force of a source.

We define this electric field in the following manner. The electric field generated by a source charge q_{source} is given by:

$$\vec{E}_{q_{\text{source}}} = \frac{kq_{\text{source}}}{r^2}\hat{r} \tag{11.3.1}$$

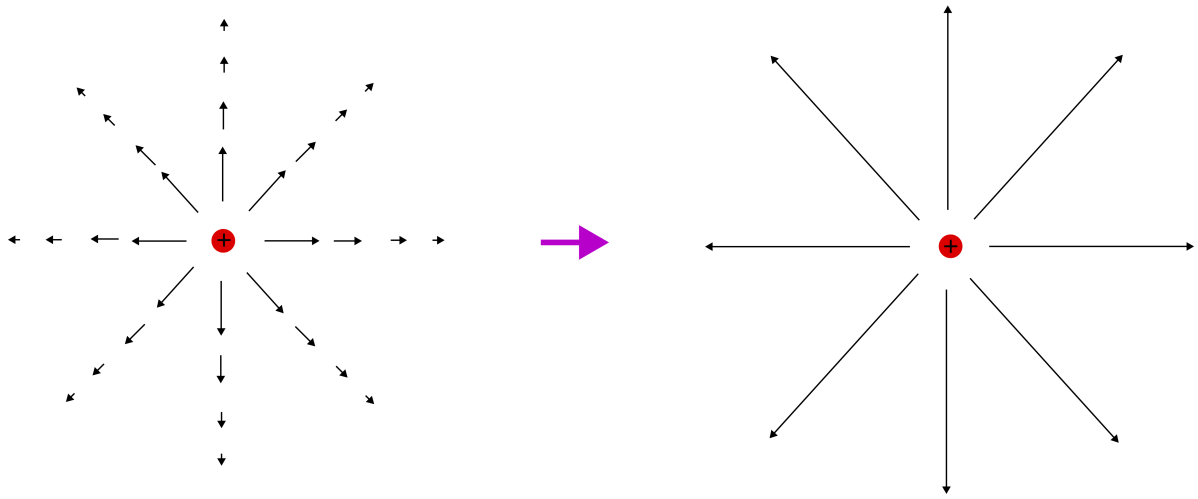
where r is the distance from the source charge to the location where the electric field is calculated. Since the radial unit vector, $\hat{r}$, points radially away from the source, if the source charge is positive, $q_{\text{source}} > 0$, then the electric field points in the same direction as $\hat{r}$, away from the source. If the charge is negative, $q_{\text{source}} < 0$, then electric field will point in the opposite direction of the radial unit vector, $-\hat{r}$, and thus, toward the source charge. Once we have established the electric field that a source charge creates, we can readily calculate the force felt by a test charge, q_{test}, which is located somewhere in the electric field generated by the source:

$$\vec{F}_{\text{on} q_{\text{test}} \text{by} q_{\text{source}}} = q_{\text{test}}\vec{E}_{q_{\text{source}}} \tag{11.3.2}$$

We can see from the above equation that the electric field has units of N/C. The above equation tells us that that if the test charge is positive, the force will point in the same direction as the electric field. For example, we have established that an electric field generated by a positive source charge will point away from it. Thus, if you placed a positive test charge in the electric field, the force it will feel will be in the direction of the electric field, away from the positive source charge. This is consistent with Coulomb's Law of two like charges repelling. Likewise, a negative source charge generates an electric field pointing toward it, so a positive test charge will feel a force in the direction of the field, or toward the negative source charge. Again, this is consistent with two unlike charges being attracted to each other. On the other hand, if the test charge is negative then the force it will feel will be in the opposite direction of the electric field. Thus, a negative test charge will feel a force toward a positive charge which generates an electric field away from it.

We would like to represent this electric field pictorially everywhere in space, as we did for multiple examples of fields in Section 11.1, in particular the vector field map of wind velocity. The electric field a vector field, so we would like to "draw a map" of the vectors around a source charge. Based on Equation 11.3.1, the electric field has a fixed magnitude for a given radial distance away from the charge, with vectors pointing away from a positive source. As the radius away from the source charge gets larger, the magnitude of the electric field decreases. Thus, in three-dimensional space there are spherical shells of vectors pointing away from the center. Although a field map is difficult to depict in 3D, we can still have a good representation by depicting the electric field in two dimensions as shown in the left picture below.

Figure 11.3.1: Electric Field Vector and Field Lines for a Positive Charge



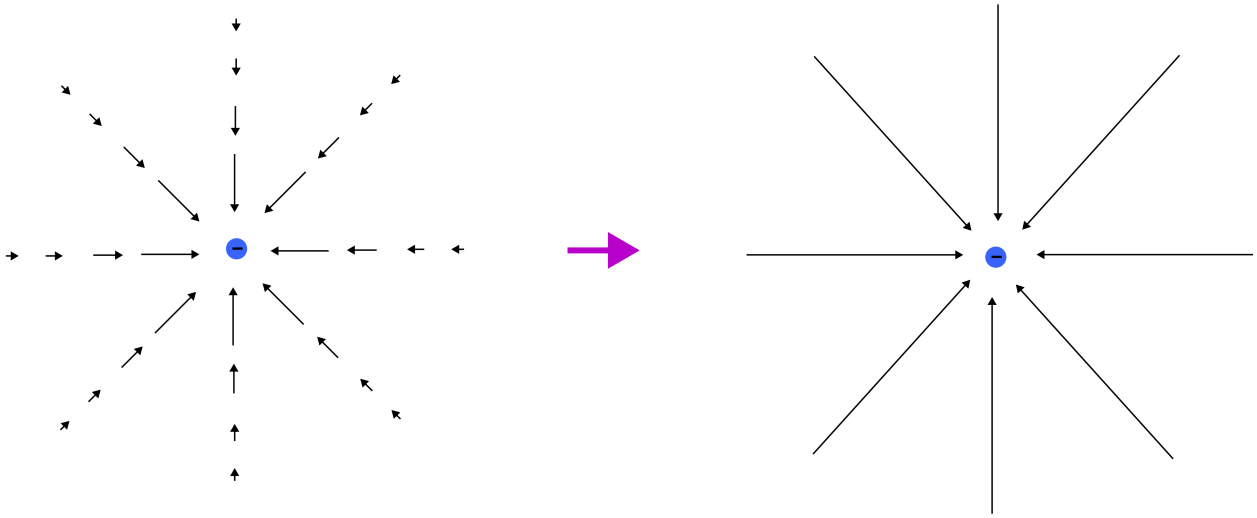
The diagram shows electric field vectors which are symmetric in magnitude at a set radius away from the source charge. They vector arrows are also getting shorter with distance, since the magnitude is decreasing. The electric field is continuous in all of space, so we arbitrary choose a few field vectors to get a good representation of the electric field. In addition, it it difficult to express the relative vector arrows in scale, since the field decreases rapidly as $1/r^2$, thus the representation above is not scaled very accurately. Instead it is often more convenient to represent the electric field using *field lines* of field vectors, as shown to the right diagram above. Field lines are "lines" that are tangent to the field vectors. Note, field lines are not the same as field vectors, but they encode some information about field vectors in the following way:

- field vectors are tangent to field lines.
- the direction of field vectors are indicated by the arrows on the field lines.
- the magnitude of the field vectors is indicated by the density of the field lines.

The density of field lines is the number of field lines per unit area. We see that the field lines are most dense near the charge and get less dense as the distance from the charge increase. This is consistent with the magnitude of the field vectors being large near the charge and decreasing as the distance from the charge increases.

Below is a similar representation for the a negative source charge. Based in Equation 11.3.1 the field vectors points in the opposite direction of $\hat{r}$, or toward the charge. The magnitude is the same as for a positive point charge, large near the charge and decreasing as $1/r^2$. The field lines representation is also shown for a negative source charge.

Figure 11.3.2: Electric Field Vector and Field Lines for a Negative Charge



The absolute density of field lines show is arbitrary and only displays the relative density with distance. So how would we distinguish the field strength for two separate charges, where one charge is double in magnitude compare to the other? In that case, if you chose eight lines to represent the field lines for the smaller charge, you would draw double the number of lines to express the

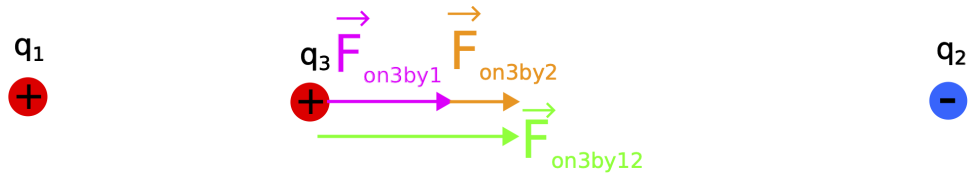
field lines for the larger charge. Thus, the number of lines drawn to represent the field of some source charge is only relevant when they are compared to other charges.

Superposition

So far we have restricted ourselves to discussing two interacting point charges, such as an electron, a proton, or an ion. In real physical situation, there are often multiple point charges interacting or a charged macroscopic object of some shape interacting with another charged macroscopic object. When this occurs, not only one but multiple charges are exerting an electric force on one particular charge. We have learned from 7B that when multiple charges act on an object, it is the net forces which affects its motion. The net force is a sum of all the forces acting on that object. It is important to remember that since force is a vector, the net force is a vector sum. Thus, to calculate the magnitude of the total force, we first need to add the forces by components and then calculate the magnitude of the total force. To review vector algebra please refer to Section 6.1 of 7B.

We will start with a simple example of a one-dimension vector addition where three charges are located along a line as shown in the figure below. We want to calculate the net force on a positive charge q_3 due to the other two charges present, a positive charge q_1 and a negative charge q_2. The force on charge 3 by charge 1 is repulsive since both charges are positive, and thus points to the right. The force by charge 2 on charge 3 is attractive, thus it also points to the right. In this case since both forces point in the same direction, you can simply add them to get the total force by both charges, $\vec{F}_{\text{on 3 by 12}}$.

Figure 11.3.3: Net Force by Multiple Point Charges



Vector addition in one-dimension is simple. If the two vectors point in the same direction, then you simply add their magnitudes to get the magnitude of the resulting vector, and the direction of that vector is the same as the direction of the two vectors which were added. If the vectors point in opposite directions, then you have to subtract the two magnitudes to get the magnitude of the resulting vector, and the direction of the resulting vector points in the direction of the vector with the larger magnitude. The sum of two vector is zero if the two vectors have the same magnitude but point in opposite direction.

In the figure above, the force by charge 1 is drawn with a bigger magnitude than the force by charge 2. If the two charges, q_1 and q_2, have the same magnitude, then this would be certainly the case since q_1 is closer in distance than q_2 to q_3. However, this does not need to be a general case in this situation, since q_2 can have a larger charge magnitude than q_1, such that the difference in charge dominates the difference in distance.

Returning to the electric field, we can describe this scenario as the two charges q_1 and q_2 being the source of the electric field, while q_3 is the test charge that experiences the force due to the presence of this (net) electric field. Thus, we can take the test charge out of the picture for now and map out the electric field at all points in space around the two electric charges. The figure below shows the electric field at a location where the charge q_3 was present in Figure 11.3.3. The electric field points away from the positive charge q_1 and toward the negative charge q_2. Thus, the net electic field points to the right.

Figure 11.3.4: Net Electric Field by Multiple Point Charges (1D vector addition)

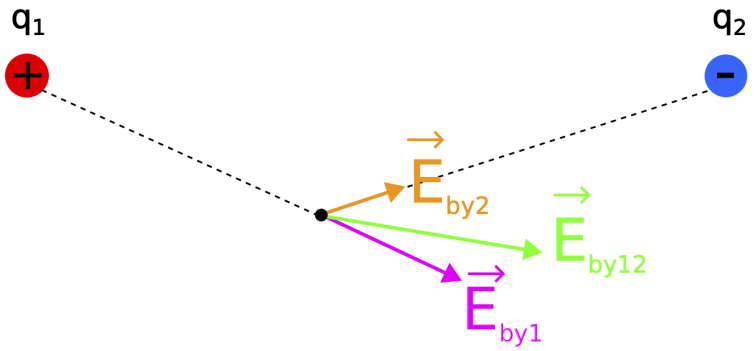
The direction of the net electric field depends on the location relative to the two charges. A second location is shown to the right of the two charges. In this case the electric field by q_1 still points to the right (away from the charge) but has a much smaller magnitude since it is further away. (The vector lengths are not drawn to scale here). The field by q_2 now points to the left (toward the charge), since the negative charge is now located to the left of this location. It has a bigger magnitude than at the location in between the two charges, since this location is closer to q_2. When adding the two vectors, we subtract the two magnitudes and since the field by q_2 is bigger than the field q_1, the net field will point to the left.

Alert

When calculating the electric field be aware of the distinction between magnitude and direction. The direction of the electric field will be dictated by two things: the sign of the charge and the location of the point where the electric field is calculated relative to the charge. The magnitude is then calculated with the absolute value of the charge in Equation 11.3.1, since its sign was already considered when determining the direction of the electric field at the specific location where it is measured.

Up to this point we have restricted ourselves to one dimension. However, the field exists everywhere in the three-dimensional space around the two charges. The figure below depicts the total electric field at a location below the line connecting the two charges. In this case since the two charges and the location where the electric field is calculated are not along a straight line, a two-dimensional vector addition is required.

Figure 11.3.5: Net Electric Field by Multiple Point Charges (2D vector addition)



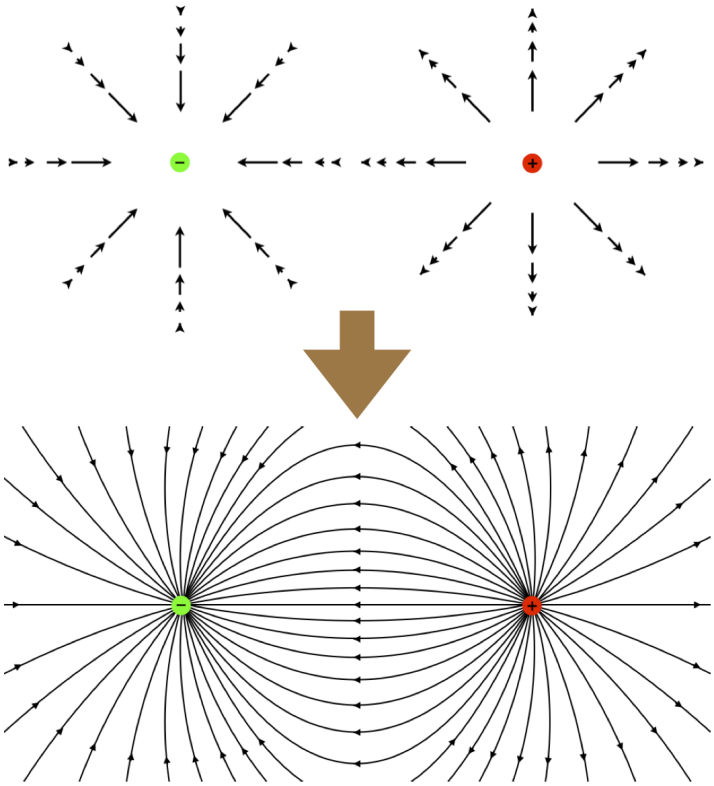
The electric field by q_1 points "southeast" away from the positive charge, and the electic field by q_2 points "northeast" toward the negative charge. Using the head-to-tail method, we can estimate the direction and the magnitude of the net electric field at that location, as shown in the figure. Mathematically, we can no longer add or subtract the magnitudes of the two electric field. Instead, we need to split each field into components, and then add the components to obtain the net field.

In general, this principle of adding the electric field due to a collection of source charges is known as *superposition*. Mathematically, the total electric field due to a collection of N source charges can be written as:

$$\vec{E}_{\text{total}} = \sum_{i}^{N} \vec{E}_i \tag{11.3.3}$$

We saw in Figures 11.3.1 and 11.3.2 that we can represent electric field vectors with fields lines. The same method applies when mapping the electic field using superposition. We first need to add all the vectors at multiple location and then construct field lines making sure that they are tangent to the net field vectors, have arrows pointing in the direction of the vectors, and have density specifying the field strength at different locations. The figure below depicts the mapping from the individual field vectors of two points charge of opposite charge to the field lines.

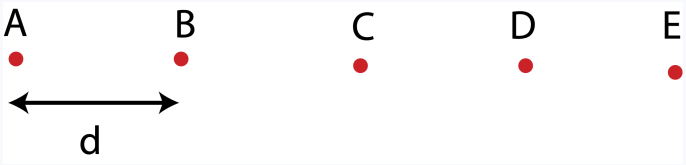
Figure 11.3.6: Electric Field Vectors and Field Lines for Two Point Charges



The field lines for this two charge source field are no longer simply straight lines, since the direction of the electric field changes as the position changes. At the location exactly halfway between the two charges the field always points to the left, since the y-components of the fields cancel. However, as the location gets closer to the negative charge for example, the direction of the field is dominated by the negative charge since its magnitude is greater. Thus, the field lines curve toward the negative charges and away from the positive ones. Also, you see the density of the field lines greater closer to the two charges, where the magnitude is greatest. At distances furthest from one of the charges, the field lines start to look a lot like those for an individual point charge, since the effect of the other charge becomes much weaker.

Example 11.3.1

Shown below are equidistant locations marked A-E separated by distance $d = 1m$. The following four charges, 1C, -1C, 2C, and -2C, are placed at locations A, B, D, and E (not necessarily in this order) such that the magnitude of the electric field at location C is maximized and points to the left.

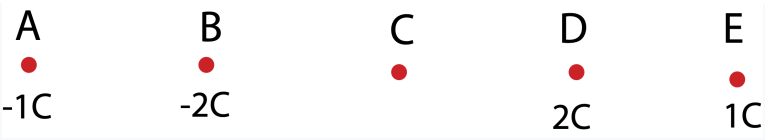


a) Determine where each charge is located.

b) Calculate the magnitude and direction of the force that a -2C charge placed at point C feels due to the presence of the other four charges.

Solution

a) To maximize electric field, the contribution from each charge should additive to the left. So, locations D and E should have positive charges and locations A and B negative ones. You also want to place the charges with the bigger magnitude closer C, since the further position the field is divided by 4 due to double distance. Below is the configuration:



b) The total electric field at location C is:

$$\vec{E}_C = \vec{E}_A + \vec{E}_B + \vec{E}_D + \vec{E}_E$$

All the electric fields point to the left at C, toward the negative charges on its left and away from the positive ones on its right. Using the equation for the electric field:

$$|\vec{E}| = \frac{kq}{r^2}$$

and plugging in the values of charges and distances we get:

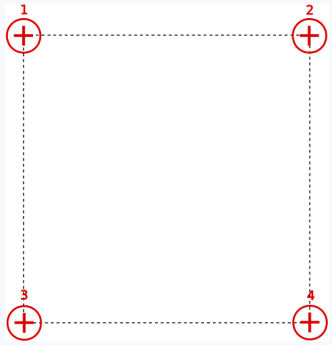
$$\vec{E}_C = -\frac{kC}{4m^2} - \frac{2kC}{m^2} - \frac{2kC}{m^2} - \frac{kC}{4m^2} = -\frac{9}{2}\frac{kC}{m^2}$$

The force on a -2C charge placed at C will be to the right, in the opposite direction of the total electric field:

$$\vec{F}_{\text{on -2C}} = (-2C)\vec{E}_C = 9\frac{kC}{m^2} = 9 \cdot 9 \times 10^9 \frac{Nm^2}{C^2}\frac{C}{m^2} = 8.1 \times 10^{10} N$$

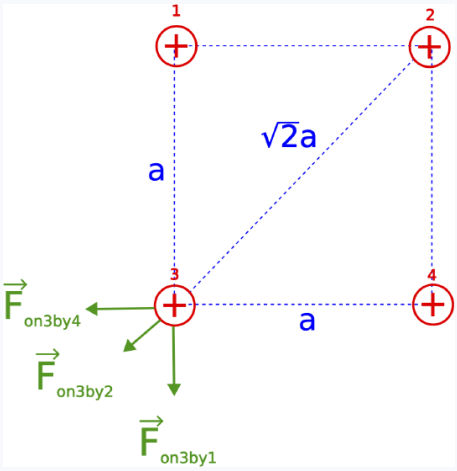
Example 11.3.2

Shown below are 4 identical positive charges located at the corners of a square. The magnitude of the force on charge 1 by charge 4 is 4.0 N. Find the magnitude of the total force on charge 3 exerted by charges 1, 2, and 4.



Solution

Let the side of the square be distance a. The relevant distances and forces on charge 3 are shown below.



Since the magnitude of the force on charge 1 by charge 4 is 4N, the same as the magnitude on charge 3 by charge 2 is also 4N, since the distance between 1 and 4 is the same as between 2 and 3, and the charges are identical, all positive with the same charge q. Thus, the magnitude of the force on 3 by 2 is given by:

$$|\vec{F}_{\text{on 3 by 2}}| = \frac{kq}{2a^2} = 4N$$

The magnitude of the force on 2 by both 1 and 4 are the same since the distance is the same:

$$|\vec{F}_{\text{on 3 by 1}}| = |\vec{F}_{\text{on 3 by 4}}| = \frac{kq}{a^2}$$

Comparing the two equations above we see that the force by 1 and 4 is double the force by 2. Therefore:

$$|\vec{F}_{\text{on 3 by 1}}| = |\vec{F}_{\text{on 3 by 4}}| = 8N$$

The direction of the force by 1 is down since the forces are repulsive. In vector form this is written as:

$$\vec{F}_{\text{on 3 by 1}} = (0, -8)N$$

The direction of the force by 4 is to the left:

$$\vec{F}_{\text{on 3 by 4}} = (-8, 0)N$$

These two vectors combined are:

$$\vec{F}_{\text{on 3 by 1}} + \vec{F}_{\text{on 3 by 4}} = (-8, -8)N$$

One way to continue is to break down $\vec{F}_{\text{on 3 by 2}}$ into components, then add to the sum of the two other forces above, and then find the magnitude. But there is a convenient shortcut when we recognize that the combined vector above will point in the same direction as $\vec{F}_{\text{on 3 by 2}}$, thus you can just add their magnitudes (it's now a 1D vector addition) to get the total magnitude of the force:

$$|\vec{F}_{net}| = \sqrt{8^2 + 8^2} + 4 = 15.3N \tag{11.3.4}$$

11.4: Conductors and Infinite Conducting Plates

Electrostatics

Before we discuss the effects of conductors on electric fields, it is essential that we make clear that we are proceeding under the following restriction: We are talking about *electrostatics*. This means that the charges present are in a state of static equilibrium – they are not moving, nor are they accelerating. We have already said that conductors allow for charges to flow freely, so how can these two things be reconciled?

If a conductor suddenly finds itself in the presence of an electric field, then the charges on that conductor will start to move as a result of the new force. As the charges move, the fields (which are affected by the placement of these charges) also change. The charges continue accelerating until the field contributions of the displaced charges cancel the external field. With zero net field, the charges no longer accelerate, and we will assume that their kinetic energy dissipates such that they also come to rest.

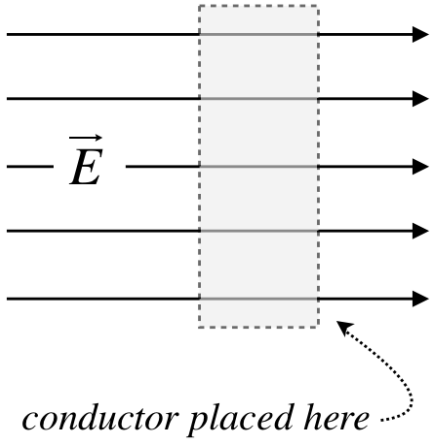
The conclusions we will draw here do not apply to the period of time when the charges are still moving around – we will only be considering the case when the charges finally reach an equilibrium electrostatic state.

Inside Conductors

We characterized electric fields as "signals" sent out by electric charges. This is of course a model (as is all of physics), and this model does not make exceptions that allow for matter to "block" this signal. The only way to affect the field at a point in space caused by one charge is to introduce a field from another charge, such that the two fields superpose.

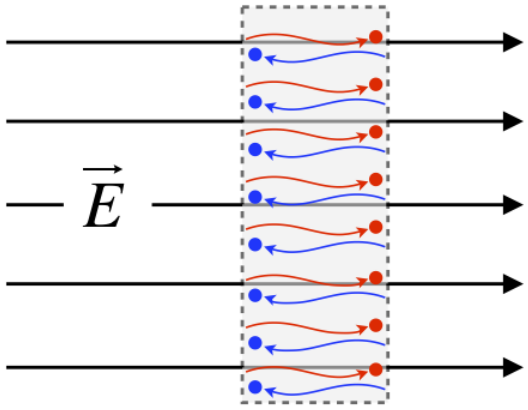
Let's now consider what happens when we suddenly introduce a uniform external electric field to a rectangular conducting slab. We represent uniform electric fields with parallel, equally-spaced field lines, and as we just said above, these field lines are not interrupted by matter, so the situation looks something like this:

Figure 11.4.1: Conductor Introduced to Field



While the conductor is neutrally-charged, it is not without electrical charge – it is made of atoms, which are comprised of protons and electrons. This charge is (by definition of the conductor) free to move. This means that the positive charges in the conductor in the diagram (depicted as red) are pushed to the right by the electric field, while the negative charges (depicted as blue) are pulled to the left.

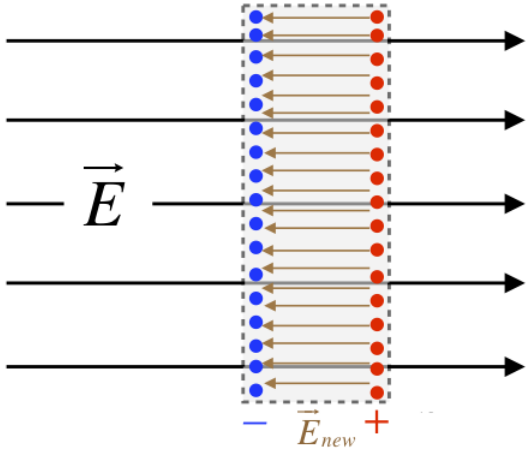
Figure 11.4.2: Charge Migrated Because of the Applied Field



[*Okay, so technically only the electrons – the negatively-charged particles – are free to move, or this conductor would not be a solid. This is a distinction that will not be at all important to us, because when negatively-charged electrons vacate a region, they leave an excess of positively-charged protons behind, which is completely equivalent to the positive charge moving into that electron-vacated region.*]

The effect of this migration of charge is the creation of two planes of charge, one on each side of the slab (we will always assume that the field is not strong enough to pull the electrons off the surface of the metal). But this separation of charge itself has consequences with regard to the field. In particular, *within the metal* a new uniform field starts to develop. This field points away from the positive charges toward the negative charges, which opposes the direction of the external field.

Figure 11.4.3: Internal Field Induced by Displayed Charges



When the field induced within the conductor is superposed with the applied field, the result is a weaker field within the conductor. The question is, how much weaker does it become? Suppose it only becomes a little weaker. That would mean that there is still some net field pointing left-to-right. But if this is the case, then more charge will migrate, making the induced field that points left even stronger, making the superposed field even weaker. In other words, we don't reach electrostatics until enough charge has moved to make the superposed internal field vanish. We therefore get the following remarkably general conclusion:

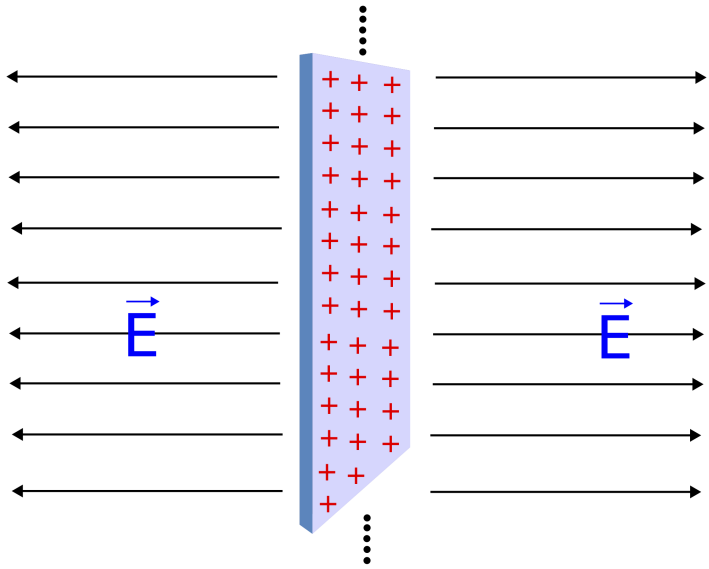
Electric fields vanish within conductors when the charges are static.

Infinite Conducting Plane

When we looked at point charges we saw that the electric field gets weaker with distance as $1/r^2$. We also explored superposition where you need to add the electric field from each point charge to obtain the net electric field. Even though the distances to each charge are no longer the same, the net field will still decrease as $1/r^2$. When the source charges are no longer discrete point charges but a continuous distrubution of charge of a specific shaped, such as a charged rod, a cylinder of a sphere, we can no longer add the field due to each discrete charge in the macroscopic charged object. These calculations are beyond the scope of this course, since sums turn into integral when you go from a discrete collection of charge to a continuous one. For some of the continuous charged distributions, the electric field on longer changes as $1/r^2$, but still decreases with distance away from the charged object.

There is one "special" charge distribution, charge equality distributed on an infinite (extremely) large conducting plane. In this case the electric field is constant around the plate, thus, it does not decrease as you move away from the plane. The electric field files for a positive infinite plane are illustrated below. If the plane was negative, the electric field lines would point toward the plane. From the field lines you can see that the electric field is constant, since the density of file lines does not charge with distance.

Figure 11.4.4: Electric Field around an Infinite Charged Conducting Plate

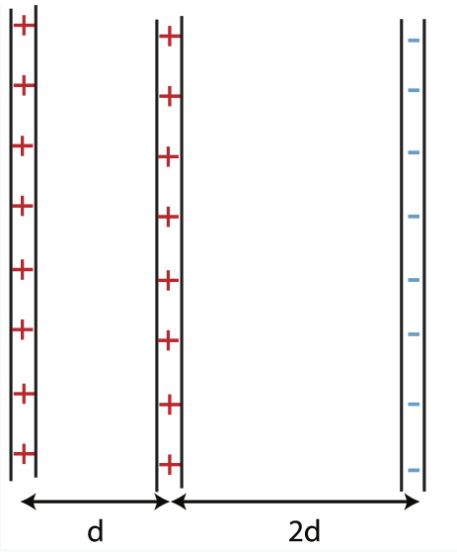


It might seem strange that the electric field is *uniform everywhere in space* for the infinite plane. Why doesn't the field get stronger closer to the plane of charge? This can be more easily seen using the field line description. If the field got stronger closer to the plane, then the field lines would have to get closer together there. But that can only happen if the field lines are not perpendicular to the plane everywhere. Due to the symmetry of the infinite plane, there is no reason to believe that the field would have any components along the surface of the plane anywhere in space. With the field only pointing along one direction, the field lines can't change their separation, and the field strength remains constant everywhere.

One might ask that although these solutions have simpler forms, how can they be "useful," as claimed above? How many infinite lines or planes of charge does one run across in real life? The answer is that *all* solutions in physics are approximations, and are only useful up to the sensitivity of the measurements. For example, if we look at the field of a finite-sized plane of charge, but look at a position in space that is very close to the plane compared to the dimensions of that plane, then treating it as "infinite" is a good approximation. What is especially nice about this approximation is that so long as we are looking at positions close to the plane compared to the distance from the edges, we don't even care what shape the plane is – it can be a circular disk, a square, or some random, jagged shape.

Example 11.4.1

Below are three infinite planes, two positively and one negatively charged with equal magnitude of charge. The separation between the planes is marked below in terms of distance d. A 4C charge which is placed equidistance between the middle positive plane and the negative plane feels a force with magnitude of 10N. Now the 4C charge is removed, and a -5C charge placed a distance d to the right of the negative plane. Determine the force (magnitude and direction) felt by a -5C charge.



Solution

The electric field is constant for infinite plates. In between the positive and negative plates, the electric field from each one of the two positive plates points to the right, away from the plane. Likewise, the field from the negative plane also and from the negative plates to the left. Since the three planes have equal charge, each one produces an electric field with the same magnitude, E. Thus, the magnitude of total electric field in between the positive and negative planes is:

$$E_{tot} = 3E$$

The magnitude of the force on a 4C charge placed in between the two plates is:

$$F = qE_{tot} = 4C \times 3E = 10N$$

Solving for the magnitude of the electric field from each plane we get:

$$E = \frac{10N}{12C} = \frac{5}{6}\frac{N}{C}$$

On the right of the plates, the electric field from the two positive planes is to the right, but from the negative plane to the left, so the total electric field is to the right:

$$E = 2E - E = E = \frac{5}{6}\frac{N}{C}$$

The force on a -5C charge will be to the left:

$$F = qE = -5C \times \frac{5}{6}\frac{N}{C} = -4.12N$$

11.5: Electrostatic Potential Energy and Potential

Electrostatic Potential Energy

Just like forces exist between two objects, the potential energy is always an energy between two objects. In Physics 7A we tied together the idea of potential energy and force. We learned that the force always points in the direction of decreasing potential energy, and the magnitude of the force depends on the rate of change (slope) of potential energy with distance. This is analogous to releasing a ball on a smooth hillside; the ball starts to roll in the direction of the steepest slope of the hill. In two or three dimensions, the force is the derivative of the potential in the direction of “steepest slope". Going back to the topographical map, contour lines with small distances between them indicate steeper hills. Objects placed here would experience higher acceleration (strongest force) down the hill than objects placed elsewhere.

It is worth repeating here the mathematical description that connects force and potential energy. Most generally, when the force is the negative of the *gradient* of the potential:

$$\vec{F} = -\vec{\nabla} PE \tag{11.5.1}$$

The *gradient operator* ∇ is short-hand for writing:

$$F_x = -\frac{dPE}{dx};\ \ F_y = -\frac{dPE}{dy};\ \ F_z = -\frac{dPE}{dz} \tag{11.5.2}$$

where x, y, and z are variables for the three spatial dimensions. We will mostly focus on potential energy between objects where the interaction is spherically symmetric (depends on separation r), such as between point charges. For these cases, Equation 11.5.1 can be written as:

$$F(r) = -\frac{dPE(r)}{dr} \tag{11.5.3}$$

where $F(r)$ is the magnitude of a force which points along the radial component $\hat{r}$. To solve for potential energy in terms of force, you can rewrite Equation 11.5.3 in terms of an integral of force over distance. For a force between two point charges described by Coulomb's Law, the electrostatic potential energy is:

$$PE(r) = -\int F(r)dr = -\int \frac{kq_1q_2}{r^2}dr = \frac{kq_1q_2}{r} + C \tag{11.5.4}$$

where the constant C is some constant, and PE has units of Joules. You can check the equation above by taking the derivative of the potential energy with respect to r and recovering the equation for the force between two charges. The constant, C, that appears in the potential energy equation should remind us that the absolute value of potential energy is arbitrary. The physically relevant quantity is the *change* in potential energy when an object moves from one location to another. For gravitational potential energy, the change in height of the object determined the change in potential energy. The actual zero value of the height, and thus potential energy, can chosen at some convenient location. For the Lennard-Jones potential we chose the potential energy to be zero when the two particles were distant and non-interacting. However, this choice of zero did not alter the is the amount of energy required to separate the two particles (break the bond), which is the *difference* in potential energy between the two particles between far apart and at their equilibrium separation.

Likewise, we can choose a convenient zero value for the electrostatic potential energy. When the two particles are far apart, then electric force becomes very weak. Thus, it makes sense to choose the potential energy to be zero when the two charges are very far apart, $r \to \infty$, as was done for the L-J potential. Taking the limit as the separation goes to infinity in Equation 11.5.4, and setting $PE(r \to \infty) = 0$, we find that $C = 0$. Thus, we write the *electrostatic potential energy* as:

$$PE(r) = \frac{kq_1q_2}{r} \tag{11.5.5}$$

Let us think about the connection of the potential energy and force to conceptually understand the equation above. If the two interacting charges are both positive or both negative, then the potential energy is positive. Moving the two charges closer together will result in a positive change in potential energy: $PE(r)$ increases as r decreases. The force points in the direction of *decreasing* potential energy, thus, moving two like charges closer together is moving them in the opposite direction of the force. On the other hand, if we start with two stationary like charges and let them move freely, they will move apart toward larger values of

r in the direction of decreasing potential energy. This is consistent with the force for like charges being repulsive and wanting to push the charges apart. You can also think about this in terms of conservation of energy, if the potential energy decreases, then kinetic energy must increase. Thus, as the two charges repel, then speed up as they separate.

If the two charges have opposite signs, then the potential energy is negative. This means that the potential energy decreases when the two charges move closer together (the potential energy will get more negative). Thus, the two charges will speed up as them come together and their potential energy decreases toward the force which is attractive for two unlike charges.

Another way to think about potential energy is the energy required to bring charges together from infinity (where potential energy is zero). If there are multiple charges, then you need to consider interactions between all the pairs involved. For example, to assemble three charges, q_1, q_2, and q_3, the potential energy is:

$$PE = \frac{kq_1q_2}{r_{12}} + \frac{kq_1q_3}{r_{13}} + \frac{kq_2q_3}{r_{23}} \tag{11.5.6}$$

where r_{12} is the separation between charges 1 and 2, r_{13} is the separation between charges 1 and 3, and so on.

Electrostatic Potential

In the previous section we defined the electric field, which is a vector field generated by a charge, a collection of source charges, or a macroscopic charged object. Once the electric field is determined, one can compute the force that a test charge feels when it is placed in the electric field. We can define an analogous field related to potential energy. Let us assign a charge or a collection of charges that generates a scalar field defined everywhere in space, which we call the *electrostatic potential, V*, sometimes referred to as simply electric potential or voltage.Then, we can analyze that amount of potential energy required to to bring a test electric charge to some location in this potential field. For one point source charge, q_{source}, the electric potential is defined as:

$$V_{q_{\text{source}}} = \frac{kq_{\text{source}}}{r} \tag{11.5.7}$$

The units of the electric potential are joules per coulomb, $\frac{J}{C}$, also known as *Volts*, *V*. The amount of potential energy required to bring a test charge, q_{test}, from infinity to a distance of r from the source charge is then given by:

$$PE = q_{\text{test}}V_{q_{\text{source}}} \tag{11.5.8}$$

Although, there is a direct relationship between force, which is a vector quantity, and energy, the potential energy is a scalar quantity which has no direction. Since the potential energy is a scalar, so is the electric potential based on equation above.

Alert

Possibly the most confusing thing to students new to electrostatics is use of the word "potential" in "electrostatic potential." This name derives from the fact that it is related to electric potential energy, but these quantities are very different, and the reader is advised to keep this in mind.

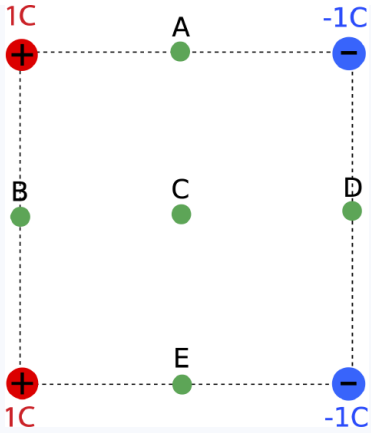
The same rules of superposition apply to the electric potential as did for the electric field. If there is a collection of charges then the total electric potential generated by these charges is the sum of the electric potentials due to each charge. Except the total electric potential is much easier to calculate since it is a scalar field, thus it is simply a sum of numbers. The only thing that you need to be careful about is the sign of each charge since the potential can be either positive or negative. For N charges, the total electric potential:

$$V_{\text{tot}} = \sum_{i=1}^{N} \frac{kq_i}{r_i} \tag{11.5.9}$$

where r_i is the distance from the location where the voltage is calculated to the charge q_i.

Example 11.5.1

The figure below shows four charges fixed at corners of a square. Points A, B, D, and E are midway between two neighboring charges and point C is at the center of the square.



a) At which of the points is the total voltage due to the 4 charges zero?

b) For the locations where the total voltage is not zero, calculate the potential energy of a -2C charge placed at those locations. Leave your answer in terms of the length of a side of the square, d, and the constant k.

Solution

a) The total voltage at each point is the sum of the voltages due to each charge. For each charge:

$$V = \frac{kq}{r}$$

The four charges have the same magnitude, so we only need to worry about their sign and distance to each location. At point B the two positive charges are closer than the negative charges, so the total voltage will be positive. At point D the two negative charges are closer than the positive charges, so the total voltage will be negative. At points A,C, and E, there is one positive charge and one negative charge the same distance away, so the voltages will cancel at those points.

b) At location B, the distance to each positive charges is:

$$r_p = \frac{d}{2}$$

We need to use Pythagorean theorem to find the distance to the negative charges:

$$r_n = \sqrt{\left(\frac{d}{2}\right)^2 + d^2} = \frac{\sqrt{5}}{2}d$$

Thus, we find that the total voltage due to the 4 charges is:

$$V_B = \frac{4kC}{d} - \frac{4kC}{\sqrt{5}d} = 2.21\frac{kC}{d}$$

At location D, the same applies except the negative charges are at the closer distance, and the positive ones are at the further one. Therefore, the total voltage has the same magnitude but opposite sign of the result above:

$$V_D = -2.21\frac{kC}{d}$$

Summary: Force, Field, Potential, and Potential Energy

In the previous section we have established the connection between electric force and field: source charges generate a vector field, and a test charge feels a force in the presence of the field. Likewise, source charges generates a potential scalar field, and a test charge placed in this field will have some potential energy. Mathematically, these relationships are given in the following way:

$$\vec{F} = q_{\text{test}}\vec{E}_{q_{\text{source}}} \tag{11.5.10}$$

$$PE = q_{\text{test}}V_{q_{\text{source}}} \tag{11.5.11}$$

The quantities on the left side of the two equations above are depend on both the source and the test charge since they describe the interaction. The quantities on the right side of the equation that multiply the test charge solely depend on the source charges.

We also used the previously learned relationship between force and potential energy, which leads directly to a new relationship between electric field and electric potential. For a point charge these are:

$$\vec{F} = -\frac{dPE}{dr}\hat{r} \tag{11.5.12}$$

$$\vec{E} = -\frac{dV}{dr}\hat{r} \tag{11.5.13}$$

The second equation comes directly from the first by dividing each side of the first equation by test charge: dividing the left side of the equation by q_{test} we get the electric field, and dividing the right side of the equation by q_{test} we get the electric potential. From the equation above we conclude that electric field points in the direction of decreasing potential, since force points in the direction of decreasing potential energy.

Here is the summary of the important concepts to keep in mind while working with these four quantities:

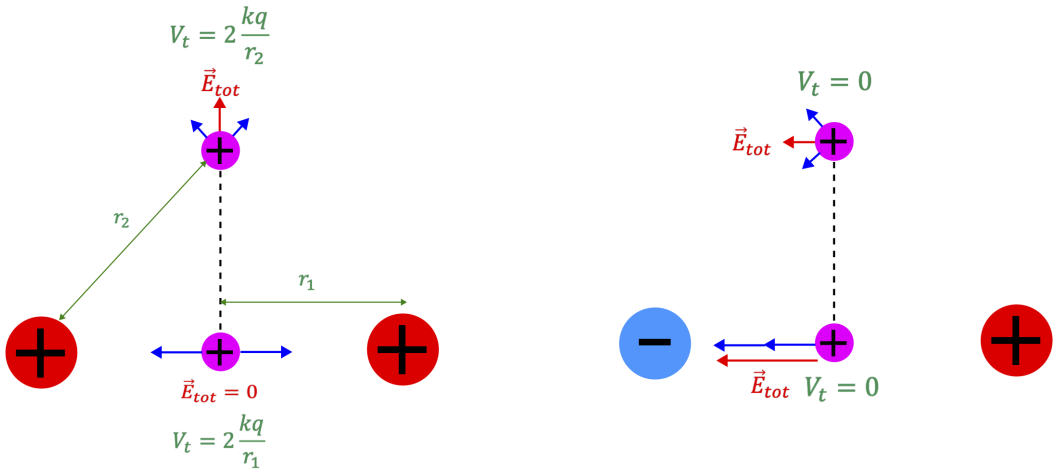
- An electric field is generated by a source charge or a collection of source charges.
- An electric field points away from a positive charge and toward a negative charge.
- For a collection of source charges, the total electric field at some location is computed by adding the electric field vectors generated by each charge.
- A positive test charge feels a force in the same direction as the electric field, and a negative test charge feels a force in the opposite direction of the electric field.
- An electric potential is generated by a source charge or a collection of source charges.
- An electric potential is always positive and decreases away from a positive charge, and is always negative and increases away from a negative charge.
- For a collection of source charges, the total electric potential at some location is computed by adding the electric potentials generated by each charge.
- A positive test charge decreases in potential energy as it moves toward decreasing electric potential, and a negative test charge increases in potential energy as it moves toward decreasing electric potential.
- Electric field points in the direction of decreasing electric potential.
- Electric force points in the direction of decreasing potential energy.

Alert

The relation between field and potential is often misunderstood, in yet another incarnation of confusing a quantity with a change in that quantity (like mistaking acceleration with velocity. Just as zero instantaneous velocity does not mean the acceleration is zero, a zero potential at a point in space does not mean that the field there is zero. Indeed, we can define the potential to be zero anywhere, no matter what the field is! It is the ***rate of change*** *of the potential that determines the field, not the* ***value*** *of the potential. It is also important to remember that we are relating a vector quantity to a scalar one. The electric potential may be constant and not changing, which only implies that the field is zero along the direction where the potential is not changing. But the field can still be non-zero since it may have components in other directions.*

The figure below demonstrates some of these common misconceptions in connecting the electric field and potential. The scenario on the left shows a positive test charge which is moved along the line in between two identical positive source charges. The situation on the right shows a positive test charge that is moved between a positive and a negative source charge of equal magnitude.

Figure 11.5.1: Field and Potential For Two Point Charges



In the scenario on the left, the force is zero on the test charge when it is right in between the two charges. The total electric field is zero at that location, since the electric fields from each source charge are equal in magnitude but opposite in direction. However, this does not imply that the potential is zero at that location. In fact, the total potential has an equal positive contribution from each charge, resulting in a total potential of $V_t = 2\frac{kq}{r_1}$. You may think how can the electric field be zero and the potential be non-zero based on their relationship given by Equation 11.5.13 The reason that the field goes to zero, is that at the central location the potential is not changing for very small displacements along the central line. Right above and right below the central location the distance to both charges is the same. Thus, the potential is constant for a small vertical displacement through the center resulting is zero rate of change of the electric potential, and thus zero electric field at the center. As the test charge moves up and away from the two charges, you can see that there is a net vertical component to the total electic field, so the charge will feel a vertical force. The electric potential decreases in that direction since $r_2 > r_1$, which is consistent with the electric field pointing in the direction of the decreasing potential.

The scenario on the right is another interesting situation, since the electric potential is zero along the line separating the two charges. This is the case since the distance to each charge is the same, but one is positive and the other is negative of equal magnitude, so the total potential adds to zero. However, this does not necessarily mean that the electric field is zero along that line. In fact as shown in the figure, there electric field is non-zero everywhere along that line and points to left, toward the negative charge and away from the positive. At first glance this does not seem to be consistent with Equation 11.5.13since the potential is not changing, but yet the field is non-zero. However, we need to remember that the field is a vector quantity. Equation 11.5.13tells us that the field must be zero along the direction where the potential is not changing. This is true for this case, since the field has no vertical component. However, the field can still be non-zero in a direction perpendicular to the central line, as seen in the figure. In fact, if you calculated the potential along the direction of the field, you would discover that it does indeed decrease in that direction since it's getting closer to the negative charge and further from the positive one.

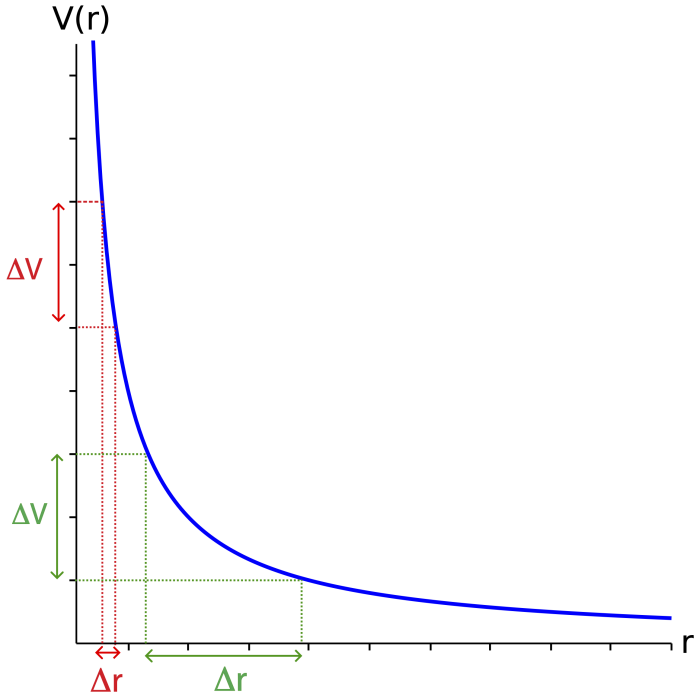
Equipotential Surfaces

In the previous section we describe a useful tool for depicting electric fields using field lines. A similar informative depiction exists for the electric potential. Going back to topographical maps in Section 11.1, we saw that it described a scalar field with lines representing location of constant value of the field (elevation). The elevation difference between neighboring lines was fixed. This means that regions where the contour lines were most dense, where regions where the slope was the largest (elevation changed most rapidly with distance). The opposite was true for regions of low density.

If we want to use a similar representation for electric potential, we will draw curves where the potential is constant, known as *equipotential surfaces*. Equation 11.5.7 tells us that the electric potential is constant along a radius, r, around a point charge. Therefore, equipotential surfaces around point charges will be spherical shells. In a two-dimensional illustration, the surfaces will become circles around the charges.

As for a topographical map where neighboring contour lines were separated by a fixed value of elevation, we want to draw equipotential surfaces separated by a fixed value of potential. Thus, we need to determine how voltage changes with distance in order to know how to space the equipotentials. The figure below is a plot of electric potential as a function of distance for a positive source charge.

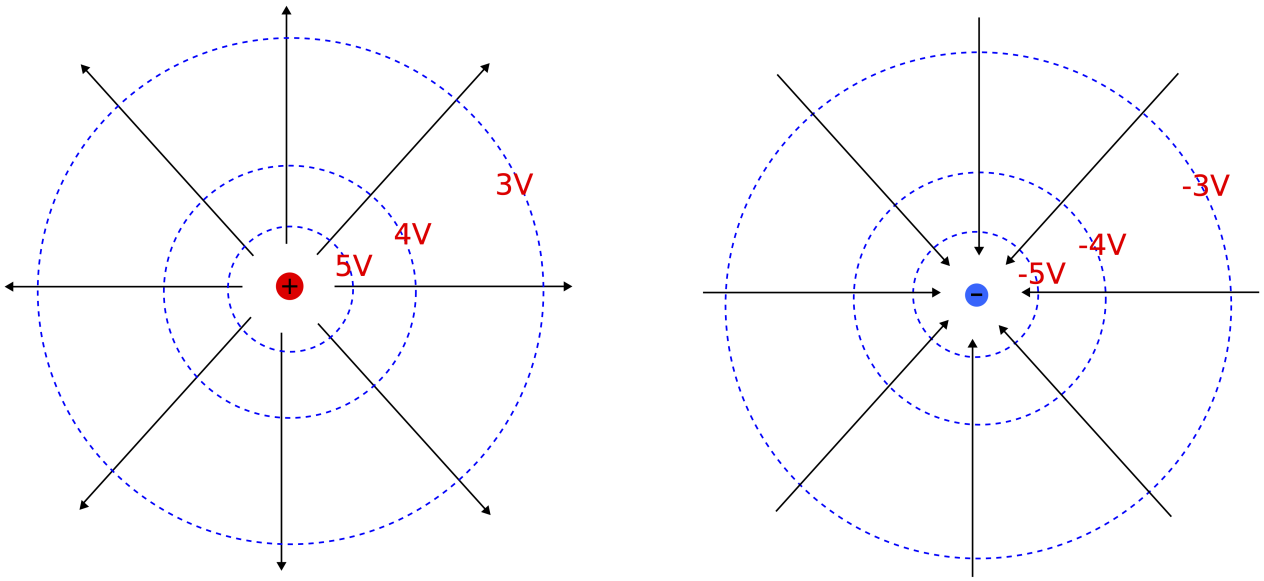
Figure 11.5.2: Electric Potential Plot for a Positive Charge



From the plot you can see that for small values of r (close to the source charge), the voltage changes more rapidly (slope is larger) compared to larger distances from the charge. Therefore, two equipotential surfaces near the charge will be separated by a smaller distance, Δr, compared to neighboring equipotential surfaces further from the charge, Δr is bigger for the same ΔV.

For a positive source charge the electric potential is positive and decreases with distance. For a negative source charge the electric potential is negative and increase with distance. All of these concepts are illustrated in the figure below, together with the field lines for a positive and a negative point charge. As shown the equipotential surfaces get less dense away from the change. The electric potential values are arbitrarily selected to change by 1V, for the sake of demonstrating the sign and change of the potential.

Figure 11.5.3: Equipotential Lines for Point Charges



The field lines appear to be perpendicular to the equipotential surfaces and point in the direction of decreasing potential. Since the electric field is the rate of change of the electric potential, when the electric potential is constant the electric field must be zero according to Equation 11.5.13 Therefore, since equipotentials describe surfaces of constant potential there cannot be any component of the electric field parallel to the equipotential. Therefore, fields lines are always perpendicular to equipotential

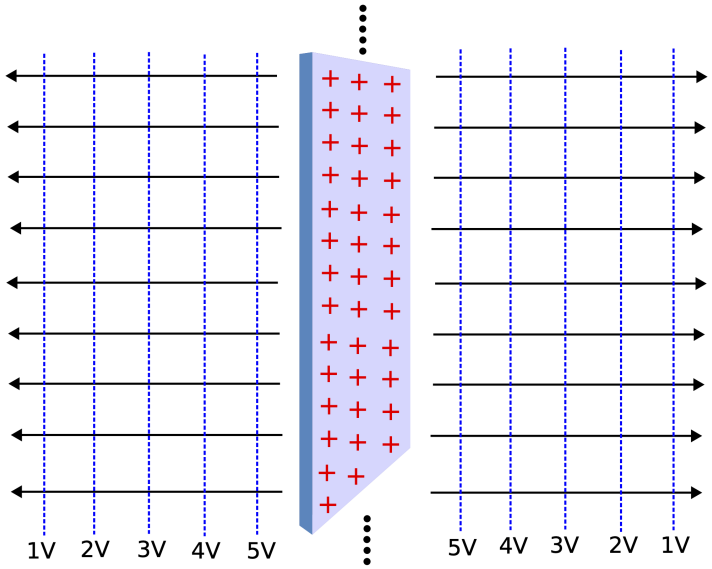
surfaces. In addition, the density of equipotential lines correlates with the density of field lines. The electric field has the greatest magnitude where the the field lines are most dense. Larger magnitude of electric field, implies larger derivative of voltage with distance according to Equation 11.5.13 Thus, the equipotential surfaces will be densest where the field lines are most dense.

In the previous section we looked as a special case of a charged infinite conducting plane, where we discovered that the electric field is constant with distance away from the plane. For this scenario Equation 11.5.13 tells us that the electric potential has to be linear with distance in order to result in a constant electric field, since the derivative of a linear function results in a constant. In other words, the change in electric potential linearly depends on the distance away from the plane:

$$\Delta V_{\text{plane}} = -E\Delta r \tag{11.5.14}$$

Therefore, the equipotential surfaces away from the plane will be equidistant, as shown in the figure below. This is consistent with the fact that the density of the electric field lines is constant, implying that the density of the equipotential surfaces must be constant as well. As previously argued, the equipotential lines are perpendicular to the field lines.

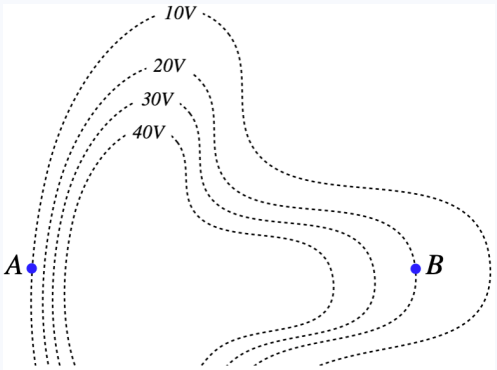
Figure 11.5.4: Equipotential Lines for an Infinite Charged Plane



Arbitrary values of the electric potential are assigned in the figure above to stress the potential is positive and decreasing away from a positively charged plane in the direction of the electric field. A positive test charge placed in this field will then will a constant force away from the plane, and its potential energy will decrease as it moves away from the plane. The opposite is true fo a negative test charge, it will feel a force toward the plane, and its potential energy will decrease toward the plane, in the direction of increasing electric potential.

Example 11.5.2

A charged particle travels through an electric field whose equipotential surfaces are shown in the diagram. The only force experienced by the charge is due to this field. The charge is moving slower at point A than it is at point B.



a. *The charge of the particle is: positive / negative / can't tell*
b. *The magnitude of the charge's acceleration is greater at: point A / point B / can't tell.*
c. *What is the direction of the charge's acceleration at each point?*

Solution

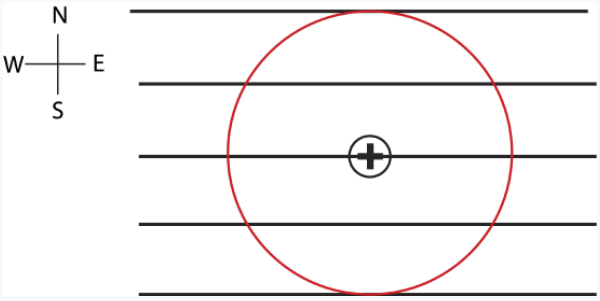
a. The particle's kinetic energy increased from point A to point B, which means that its potential energy went down. But its electrostatic potential went up, so since $\Delta U = q\Delta V$, then $\Delta U < 0$ and $\Delta V > 0$ means that $q < 0$.

b. The equipotentials all differ by equal voltages, so those that are closer together indicate a region where the electric field is stronger. The field is therefore stronger at point A, which means it experiences a greater net force there than it does at point B.

c. The force due to the electric field must be parallel to the electric field, which must be perpendicular to the equipotential surface. So the forces at points A and B must be either to the left or to the right, but can we tell which way? The field points from higher potential to lower potential, so at point A it points left, and at point B is points right. The charge is negative, so the forces are opposite to the electric field directions. The particle accelerates to the right at point A and to the left at point B.

Example 11.5.3

Shown below are equidistant equipotential horizontal lines due to some source incremented by 4V. The top equipotential line represents 0V. The bottom equipotential line is 6 meters below the top equipotential line and has a lower voltage. Also, located on one of the lines and at the center of a circle is a 2×10^{-9} C positive charge. The circle represents locations where test charges are placed in order to determine the forces they would feel.



a) Calculate the force that a negative charge with magnitude of 5×10^{-9} C feels placed on the circle directly east from the charge in the center. Express the force in terms of magnitude and direction.

b) Calculate the total voltage at the location of the added charge in part a).

c) Calculate the change in potential energy of the negative charge from a) when it moves from directly east to directly south on the circle.

Solution

a) In order to find the net force on the negative test charge, we need to consider the total electric field at the location of the charge due to the source that generated the equipotential lines and due to the positive point charge at the center of the circle.

The electric field due to the equipotential lines is constant since the lines are equally spaced. Its magnitude is:

$$E_{\text{lines}} = \frac{\Delta V}{\Delta r} = \frac{4 \times 4V}{6m} = 2.67 \frac{V}{m}$$

The electric field points in the direction of decreasing voltage, thus, it points South:

$$\vec{E}_{\text{lines}} = (0, -2.67) \frac{N}{m}$$

The radius of the circle is 3 m. Thus, the magnitude of electric field 3 m from the positive point charge in the center is:

$$E_q = \frac{kq}{r^2} = \frac{9 \times 10^9 \frac{Nm^2}{C^2} \cdot 2 \times 10^{-9} C}{(3m)^2} = 2\frac{V}{C}$$

The electric field due to the central charge points away from it since it's positive. Thus, it points East at the location of this charge:

$$\vec{E}_q = (2, 0)\frac{V}{m}$$

The total electric field is then the sum of the two electric fields above:

$$\vec{E}_{tot} = \vec{E}_{\text{lines}} + \vec{E}_q = (2, -2.67)\frac{V}{m}$$

The force on the negative charge is:

$$\vec{F} = q\vec{E}_{tot} = (-5 \times 10^{-9} C)(2, -2.67)\frac{V}{m} = (-10 \times 10^{-9}, 13.35 \times 10^{-9})N$$

The magnitude of the force is:

$$|\vec{F}| = \sqrt{(10 \times 10^{-9})^2 + (13.35 \times 10^{-9})^2} = 1.67 \times 10^{-8} N$$

The direction is north of west at an angle of:

$$\theta = \tan^{-1}\frac{13.35}{10} = 53.16^\circ$$

b) From the equipotential lines, the voltage of the third line from the top is:

$$V_{\text{lines}} = -2 \times 4V = -8V$$

From the charge in the center:

$$V_q = \frac{kq}{r} = \frac{9 \times 10^9 \frac{Nm^2}{C^2} \cdot 2 \times 10^{-9} C}{3m} = 6V$$

The total voltage at that location is the sum of the two voltages:

$$V_{tot} = V_{\text{lines}} + V_q = -8V + 6V = -2V$$

c) Since the circle defines an equipotential for the central charge, only contribution to the change of the potential energy are the equipotential lines:

$$\Delta PE = q\Delta V_{\text{lines}} = (-5 \times 10^{-9} C)(-8V) = 40 \times 10^{-9} J$$

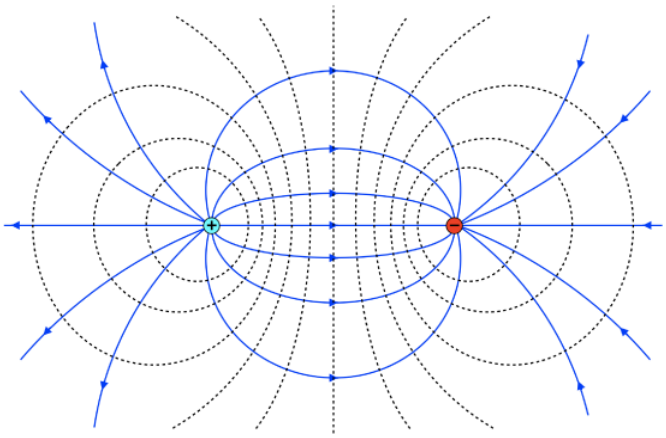
11.6: Electric Dipole

Definition

We know well how to deal with single point charges and an infinite charged plan which generates a simple form of an electric field, but of course physical systems rarely behave in such a simple way. What we are going to look at here is a model for *two* point charges that are equal in magnitude and opposite in sign. The reason this is such an important package to develop is that it appears so much in nature, in the form of neutrally-charged molecules.

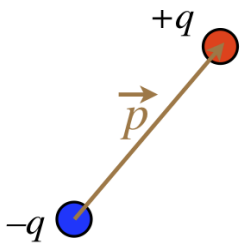
Consider two equal point charges, one positive, and the other negative, that are held rigidly at a fixed separation distance (if you like, you can imagine a tiny rigid rod holding them at fixed relative positions). We have already seen what the field of such a dipole configuration looks like, in Figure 11.3.6. The figure below shows the field lines with equipotential lines included. Since the field lines are now curved, the equipotential circles are also distorted such that they stay perpendicular to the field lines. The density of both the field and the equipotentials is highest near and between the charges, where the additive effect fo the electric fields is greatest with both fields pointing in the same direction.

Figure 11.6.1: Equipotential Surfaces of a Dipole



We could forever treat such a configuration as a combination of two point particles, but it is helpful to package them so that we can treat them as a single entity and not have to go back and recalculate things. To that end, we define a vector quantity known as an *electric dipole moment*, $\vec{p}$, as shown in the figure below.

Figure 11.6.2: Electric Dipole Moment



The magnitude of the dipole moment is defined as the product of the absolute value of one of the two charges, multiplied by the distance separating the two charges:

$$\left|\vec{p}\right| \equiv q\, d \tag{11.6.1}$$

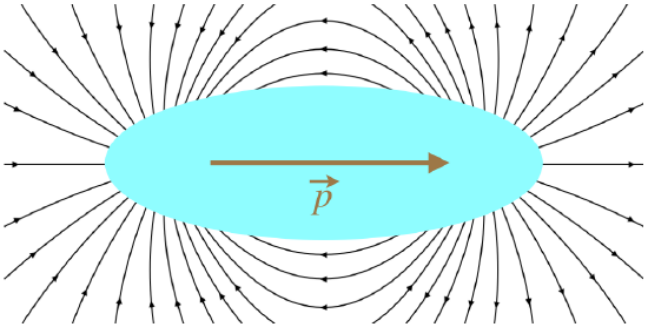
The direction of the dipole moment is that it points *from the negative charge to the positive charge*.

Alert

Chemists typically define the dipole moment as pointing in the opposite direction. When creating a "package" for later use, how you define it is up to you. We will see that there are compelling reasons (at least in physics applications) for defining it as above.

Note that the dipole moment is not the same as the dipole electric field. It may seem funny to even mention this, as these two quantities are not even close to being the same, but it does come up. One place where it gets confusing is that the dipole moment points in the *opposite* direction as the electric field between the two charges. But as we are forming a package with these two charges, what happens *between* them is of no consequence. When it comes to the direction of the *dipole field*, the dipole moment direction makes perfect sense.

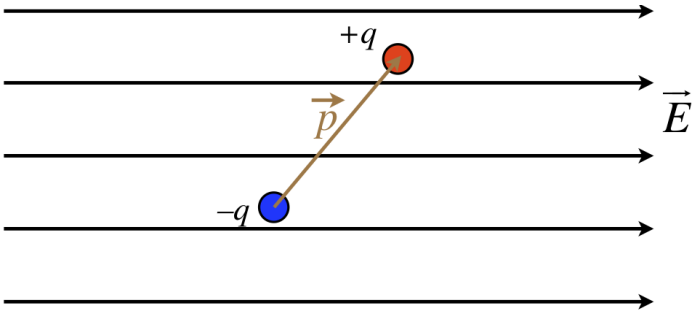
Figure 11.6.3: Field of a Dipole



Potential Energy

We consider now the effect that a uniform electric field has on a dipole. Note that while we will be assuming a uniform field, in reality we mean that the amount that the external field changes across the length of the dipole is negligible. Also, as will generally be the case going forward, when we draw a diagram of a uniform field, we will represent it with a set of parallel field lines.

Figure 11.6.4: Dipole in a Uniform Field



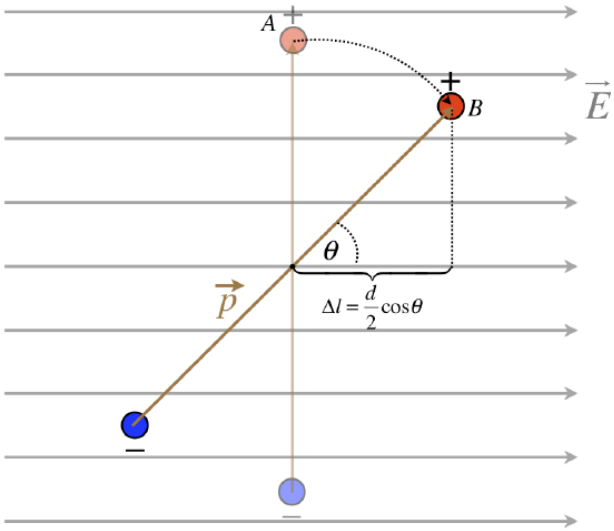
We begin by considering the force on the dipole. Certainly each individual charge feels a new force from the field, but the charges are equal in magnitude, and the forces act in opposite directions, so the net force on it is zero.

Alert

If the field is not uniform, then the dipole can experience a net force! This might seem odd, given that the "package" of two charges is neutrally-charged, but it is an important physical effect to be aware of (we will discuss it in more detail later).

With no net force, the center of mass of the dipole will not accelerate, but there will clearly be a *torque* exerted on this object with will allow the dipole to rotate about its center of mass. If there is a net torque exerted, there should be a potential energy associated with this interaction. Suppose we release the dipole in the diagram above from rest. Clearly it will begin to rotate, which means it will gain kinetic energy thus some potential energy must be lost.

Figure 11.6.5: Potential Energy Change for a Rotating Dipole in a Uniform Field



Let us consider the change in potential energy of rotating the dipole by some angle θ as shown in the figure above. The horizontal displacement due to this rotation is calculated as Δl in the figure. We have calculated in the previous section that the potential changes linearly in a constant electric field. For a positive charge in the dipole the change in potential energy is then:

$$\Delta PE_p = q\Delta V = -qE\Delta r = -qE\frac{d}{2}\cos\theta \tag{11.6.2}$$

For the negative change, the horizontal displacement has the same magnitude but a negative sign since the charge is moved in the direction of increasing potential:

$$\Delta PE_n = -q\Delta V = qE\Delta r = -qE\frac{d}{2}\cos\theta \tag{11.6.3}$$

Combing the two results above, we find that the total change in potential energy is:

$$\Delta PE_p = -(qd)E\cos\theta \tag{11.6.4}$$

Note that the energy is a minimum when the dipole moment aligns with the external electric field.

11.7: Magnetic Field

Field of a Magnetic Dipole

Everyone has at least a passing experience with magnets – little metal disks we stick to refrigerators to hold papers we need to remember. From our experience, we know that if we put two magnets together a certain way, they stick together, and if we turn one of them around, they repel. So they clearly have a directionality to them. 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 12^{th} century) makes use of the interaction between bar magnets and the Earth (which, for the purposes of magnetism, is one large bar magnet). 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.

Since there are two opposite poles on a magnet, the closest analogy in electricity is a dipole which has two opposite charges. Indeed, if we put two dipoles end-to-end one way, they will attract, and if we turn one of them around, they will repel.

Figure 11.7.1: Attraction of Aligned Electric Dipoles

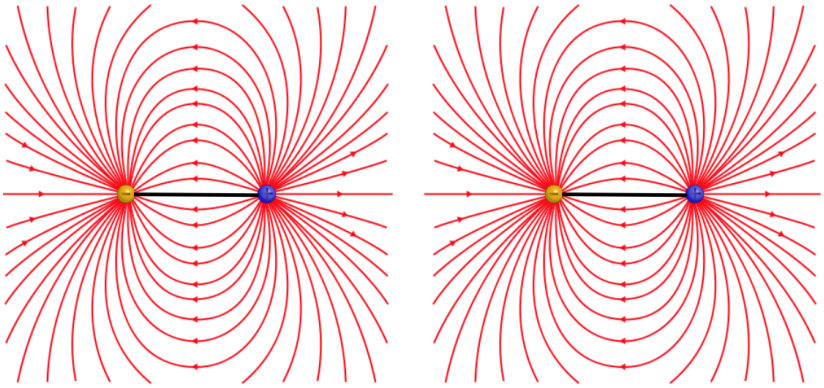
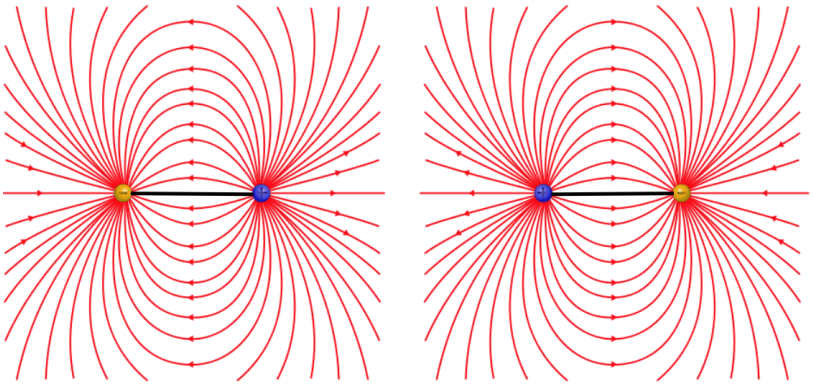


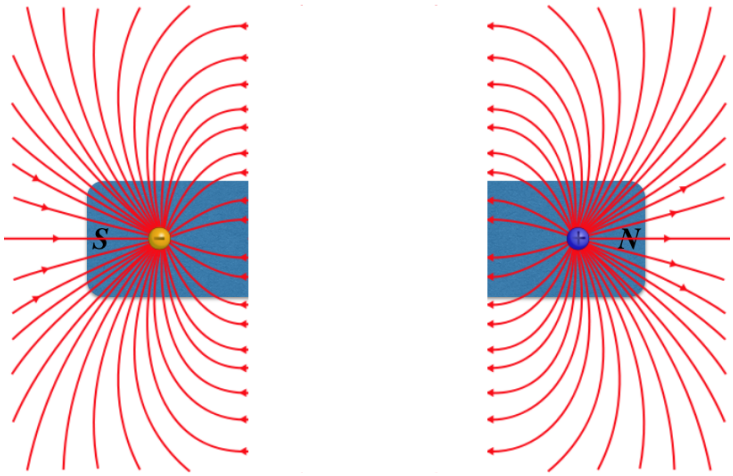
Figure 11.7.2: Repulsion of Anti-Aligned Electric Dipoles



The attraction and repulsion occur because the there is a field created by one dipole that points in the direction outward from the positive charge, and the field gets weaker with distance, so the other dipole will feel a net force according to whichever of the two charges is closer to the dipole creating the field. In magnetism, we call the end of the magnet from which the outward-going field lines emerge, the *north pole*, and the end into which the field lines converge, the *south pole*. From the figures above, it's clear that the dipoles whenever like poles are brought together, and attract when opposite poles are brought together.

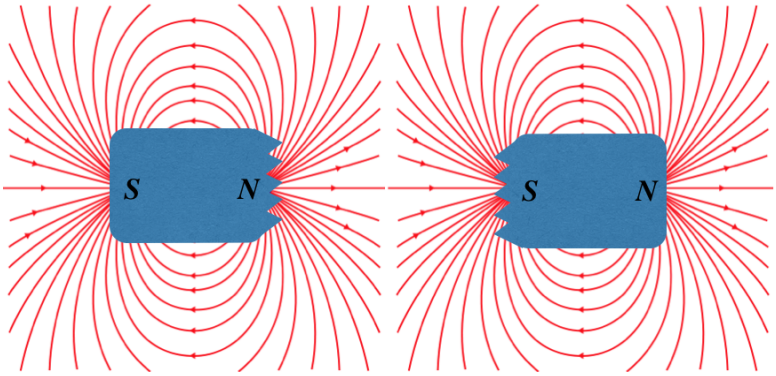
Okay, so this looks like a reasonable explanation for how magnets work, so if we want to isolate the two individual magnetic charges (a "north charge" and a "south charge"), all we have to do is cut the magnet in half, right?

Figure 11.7.3: Isolating Magnetic Charges from a Magnet – An Attempt



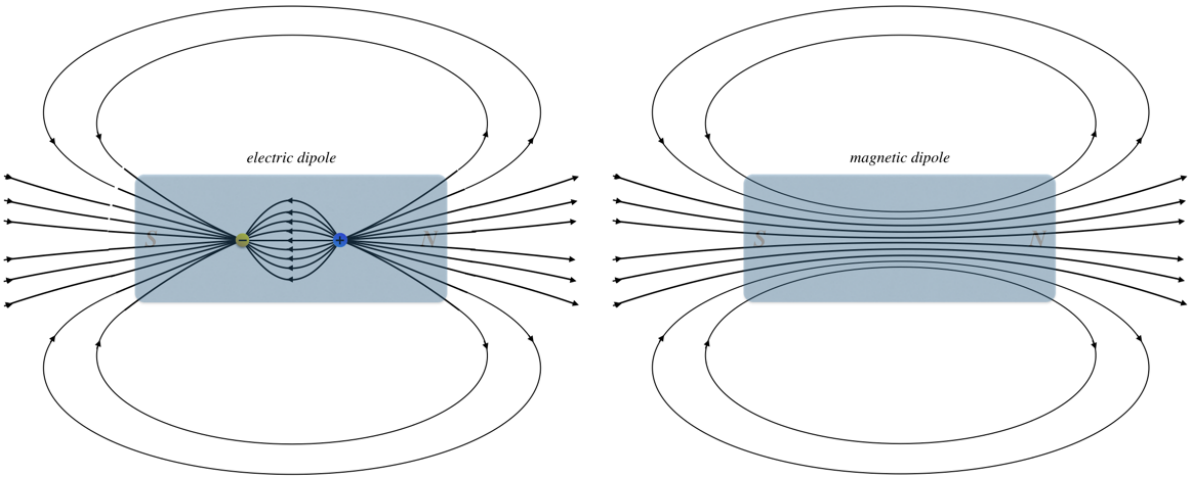
Well, try as we might, ***this never happens***. Instead, every time we cut a magnet with two poles into two pieces, we just get two more magnets with two poles!

Figure 11.7.4: Isolating Magnetic Charges from a Magnet – What Actually Happens



If we examine the the field lines for a bar magnet closely and compare them to an electric dipole field, we see how fundamentally-different the two fields are. For the electric dipole, the field changes direction between the two poles, while for the magnetic case, the field lines continue straight through:

Figure 11.7.5: Comparing Field Lines of Electric and Magnetic Dipoles



Outside the dipoles, the fields look the same, but they are clearly different, which we can characterize in the following way:

Magnetic field lines always form closed loops, while electric field lines begin and end on electric charges.

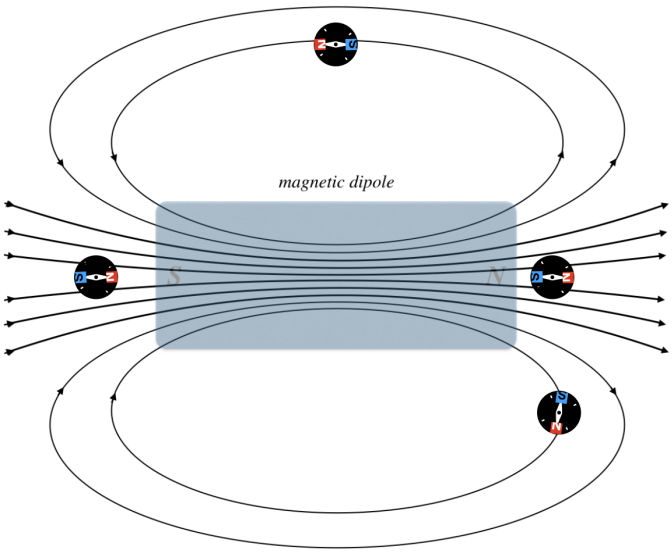
Put another way, unlike electric fields which form their dipole fields from two *monopoles*, there don't seem to be any magnetic monopoles. Or at least we have never been able to detect a magnetic monopole, despite many decades of experimental search for them. It turns out that electromagnetic theory doesn't exclude the possibility of the existence of these point charges of magnetism, but ultimately our theories have to agree with what we observe in experiments, so at least for the moment (and for the duration of this class), we will maintain the position that they simply don't exist.

Magnetic Dipole

For the electric field we described the force that a electric charge feels in an external field. Then, we also explored the force that an electric dipole feels in the presence of an external field in Section 11.5. Let us compare the electric dipole interaction with one for a magnetic dipole, since there is no magnetic monopole to make the comparison with a point charge. All magnets are magnetic dipoles, so we want to explore what happens when you place a magnet in an external magnetic field which, for example, is generated by another bar magnet. The "test" dipole magnet can be a magnetic needle of a compass.

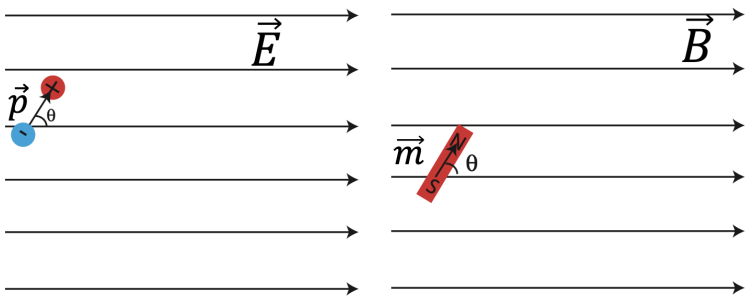
The bar magnet produces a magnetic field in its surrounding space, and the magnetic needles of the compass reorient themselves because the field exerts a force on them. For example, the compass needle that is exactly to the right of a bar magnet is closest to the North pole of the bar magnet. Therefore 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 as shown in the diagram below. The opposite happens at the south end of the magnet. Placing the compass at different locations around the magnet, we discover that the magnetic needle always lines up with the external magnetic field, such that the north end of the needle points in the direction of $\vec{B}$.

Figure 11.7.6: Magnetic Dipole in an External Magnetic Field



For electric dipoles we defined the electric dipole moment, $\vec{p}$, and showed that it lines up with the external electric field, $\vec{E}$. The potential energy of this interaction was described by the equation shown in the figure below. Likewise, let us define the *magnetic dipole moment*, $\vec{\mu}$, which points from the south end to the north end of a magnet. Thus, magnetic dipole moment, $\vec{\mu}$, will line up with the external magnetic field, $\vec{B}$. The equation that describes the magnetic potential energy of this interaction is shown below.

Figure 11.7.7: Comparing Electric and Magnetic Dipole Interactions



$$PE_{electric} = -|\vec{p}||\vec{E}|\cos\theta \qquad PE_{magnetic} = -|\vec{\mu}||\vec{B}|\cos\theta$$

Although, the two dipoles describe physically different situations, the analog between the two interactions is strikingly similar.

11.8: Magnetic Force

Forces on Moving Charged Particles

If we run currents through two parallel wires, something unexpected happens – the wires exert forces on each other! One might be inclined at first to explain this by claiming that putting currents through the wires puts electric charge into them, and that the electric charges are exerting electrical forces on each other. But this is not correct. Current is simply a steady flow of charge – there isn't any charge built-up in the wire. For every electron that enters the wire, a corresponding electron exits it. So the force must have something to do with the *motion* of the charges. After further experimentation, we find that if we place a stationary net charge near a conducting wire with current, there is no force between them. So apparently there must be motion for *both* sets of charges in order to exhibit this force. These observations is what eventually led to the connection between electricity and magnetism.

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. The ideas of how a magnetic field affects moving charges were not known until the mid-1800s. Before that, the only thing known about magnetism was that some materials can produce magnetic fields and these attract (or repel) certain kinds 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, known as *electromagnetism*. 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.

While this force involves electric charge, it clearly is not electrical in nature. That is, it is altogether different from the Coulomb force. We therefore give it a different name. We call it the *magnetic force*. Like the electric force, we will explain it in terms of a vector field. And as with the electric force, this will require the two-step theory of first explaining how a charge acts as a source for the field, and then how another charge reacts to being in the presence of a field. We are going to hold off on the first step for now, and focus on the second step, which means we will just start with a magnetic field (without worrying about how it got there beside what we already know about magnets from the previous section), and discuss how the field exerts a force on the moving charge.

We will approach this topic as if we were performing experiments to extract the information we want. Here is a list of our observations from these experiments:

- *The strength of the magnetic force on a charge is proportional to the magnetic field through which the charge is moving* – This is not surprising, as it was also true for the electric force and field, but more to the point, it really is a result of our *definition* of magnetic field.

$$\left|\vec{F}_B\right| \propto \left|\vec{B}\right| \tag{11.8.1}$$

[*Note: The traditional variable for magnetic field is B.]*

- *The strength of the magnetic force on a charge is proportional to the magnitude of the charge* – Again, not surprising.

$$\left|\vec{F}_B\right| \propto q\left|\vec{B}\right| \tag{11.8.2}$$

- *The strength of the magnetic force on a charge is proportional to the speed at which the charge is moving though the field* – This was mentioned above, and it is the first divergence form the electrical case.

$$\left|\vec{F}_B\right| \propto q\left|\vec{v}\right|\left|\vec{B}\right| \tag{11.8.3}$$

- *The strength of the magnetic force on a charge varies depending upon the relative directions of the magnetic field and the charge's velocity vector* – Now *this* is new! Specifically, we find that the force is zero if the charge happens to be moving parallel to the field, and is its strongest when the field and velocity are perpendicular to each other. Further experimentation reveals that the strength of the force is proportional to the sine of the angle between the field and velocity vectors. Thus, we obtain the *magnitude* of the *magnetic force* on a moving charge in an external field:

$$\left|\vec{F}_B\right| = |q|\left|\vec{v}\right|\left|\vec{B}\right|\sin\theta \tag{11.8.4}$$

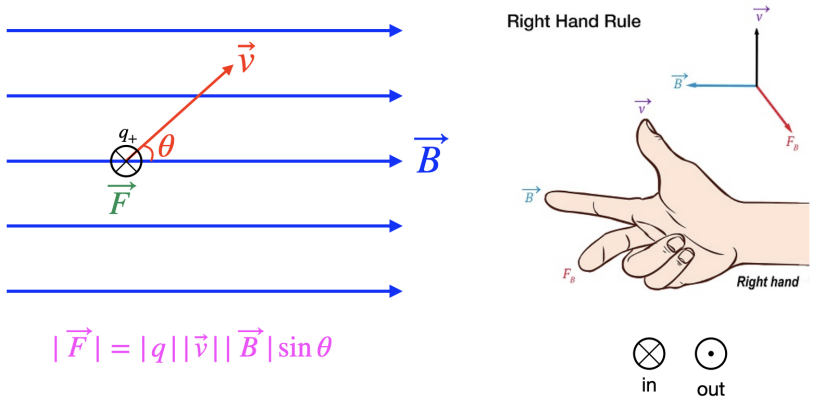
- *The direction of the magnetic force is perpendicular to both the direction of the velocity and the direction of the magnetic field* – This is quite different from the electric case, for which the direction of the force and the field are always in the same or opposite directions.

When combining the magnitude and the direction of the magnetic force, mathematically it is written using the cross product:

$$\vec{F}_B = q\vec{v} \times \vec{B} \tag{11.8.5}$$

Since we do not calculate cross-products directly in this course, we can instead determine the magnetic force vector by using using Equation 11.8.4to calculate its magnitude, and using a *right-hand-rule for the direction of the magnetic force* depicted in the figure below. To use the right-hand (RHR) rule, as the name implies use your *right* hand, point your thumb in the direction of the charge's velocity, your index finger in the direction of the component of the magnetic field which is perpendicular to the velocity, and your middle figure which is perpendicular to the other two will tell you the direction of the force. It is also important to note that like the case of the electric force, a negatively-charged particle experiences a magnetic force in the opposite direction of a positively-charged particle under the same conditions. Thus, if the charge is negative, you can still apply the same right-hand-rule but flip the result by 180^o to obtain the correct direction of the force on a negative charge. The figure below shows the relationship of the velocity, field, and force vectors in an example. It should be noted that the symbols $\odot$ and $\otimes$ will start playing common roles in diagrams from here on, and they represent the directions "out of the page" and "into the page", respectively.

Figure 11.8.1: Magnetic Force Direction and Magnitude



If there is more than one source of magnetic field, we apply the same rules of superposition as we did for electric fields by adding the magnetic field vectors generated by the individual sources to obtain the net field. In addition, there's no reason that electric and magnetic fields can't coexist in the same space. When they do, the force on the point charge is the vector sum of the two forces. The combination (often referred to as the *Lorentz force*) is therefore:

$$\vec{F} = q\left(\vec{E} + \vec{v} \times \vec{B}\right) \tag{11.8.6}$$

Before moving on, we should say a word about units. For the equation given above to have proper units, the untis for magnetic field must be force divided by charge-times-velocity. Given how many different names we have for various units, we can of course express this many ways, but once again we will give this quantity its own name:

$$[B] = \frac{[F]}{[q]\,[v]} = \frac{N}{C \cdot \frac{m}{s}} = \frac{N}{A \cdot m} \equiv T \quad (\text{“}Tesla\text{”}) \tag{11.8.7}$$

It turns out that a magnetic field with a magnitude of 1 T is quite strong. It is not beyond common experience (neodymium bar magnets get to this strength), but more commonly encountered magnetic fields (such as that of the Earth, or a compass needle) are significantly less, and it is more common to see these field strengths described in units of *Gauss* (*G*), which the simply one ten-thousandth of a Tesla. The strength of the Earth magnetic field near its surface ranges from 0.25 to 0.65 G.

Field From a Moving Point Charge

When we first started discussing magnetism, we noted a force between two current-carrying wires. From there, we focused on the fact that a magnetic field affects only *moving* electric charges, but it should be equally clear that the *source* of a magnetic field must also be moving electric charges. One might object that we just said that magnetic fields don't have point sources, so what difference does it make that we insist that the point source be moving? We will see that this makes all the difference, because this leads to a field that doesn't point directly toward or away from that charge – the direction of the field is determined by the direction of the velocity vector.

As different as the magnetic field is from the electric field, there are still so many striking similarities that it is useful to describe the features of the magnetic field from a moving point charge in parallel with the Coulomb electric field. This magnetic analog of the Coulomb field is called the *law of Biot & Savart*.

- In the Coulomb case, we started with the fact that the field strength is proportional to the magnitude of the charge emitting the field. In the magnetic case, the field strength is also proportional to the magnitude of the charge, but since the charge must also be moving, it turns out that the field strength is also proportional to the charge's *speed*. This agrees with the observation that there is no magnetic field if the charge is stationary.
- Next we consider how the strength of the field weakens with distance from the source. In this case, the two fields behave identically – with an inverse-square law.
- The direction is where these two fields differ the most. In the Coulomb case, the field points directly toward or away from the point charge. Put another way, if the source charge is at the origin, then the electric field at a position in space described by a position vector $\vec{r}$ points in a direction that is parallel to that position vector. The magnetic field, by contrast points *perpendicular* to that position vector. This doesn't narrow down the direction, however, as there is an entire plane that is perpendicular to this vector. This is where the velocity vector direction comes in – the magnetic field is also perpendicular to the direction defined by the velocity vector. We already know a way of expressing a vector that is perpendicular to two other vectors at the same time – it must be parallel to the cross product of those two vectors.

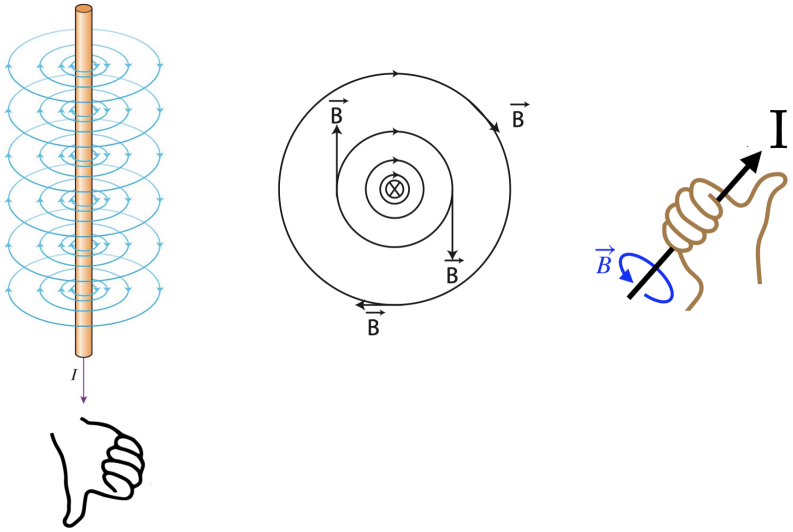
Field From a Current in a Straight Wire

It is far more common to have physical situations where a magnetic field is created by a current-carrying wire rather than by a moving point charge. 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 is its magnitude, where it points (recall it is a vector), and how does it vary with position. In other words, we want to map the magnetic field around the 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 (running up and down through the table surface). Let's assume the current direction to be coming from the top and going toward the bottom of the table, as drawn on the left in the figure below. When there is no current in the wire, all needles point in the same direction (magnetic north). As soon as 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 recognizable pattern: if we draw a circle on the table with the wire at the center and we place the compasses along the circle, we will notice that all the compass needles will orient themselves *tangential to the circle*. In other words, the field lines for the magnetic field, $\vec{B}$, generated by 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 down into the table, we find that the magnetic field direction clockwise along the circles if viewed from the top. The figure below has both the side an the top view (with current going into the page) of the wire. If we flip the direction of the current, the compass needles will still point tangent to the circle, but now their north poles point in a counterclockwise direction.

Figure 11.8.2: Direction of Magnetic Field Around a Current Carrying Wire



Thus, there is yet another very convenient *right-hand-rule for the direction of magnetic field around current carrying wires*. It is demonstrated in the figure above. You need to point the thumb of your right hand in the direction of current. Then the curling of your fingers will indicate the direction of the magnetic field circles that are formed around the wire.

Let's mathematically relate every quantity that we have talked about. We expect that all points at the same distance r will have the same magnitude, since the compass needles around a wire arranged in a circle with radius r. It is also not unexpected that the magnitude of the field will be larger if we have a larger current, I. It turns out that the field is proportional to the current. For mathematical reasons we won't go into, when you convert point charges to currents the relationship between magnetic field and distance no longer follows the inverse-square law, but is now proportional to the inverse of distance. Precisely, the magnitude of the magnetic field around a current carrying wire is given by:

$$|\vec{B}| = \frac{\mu_o I}{2\pi r} \tag{11.8.8}$$

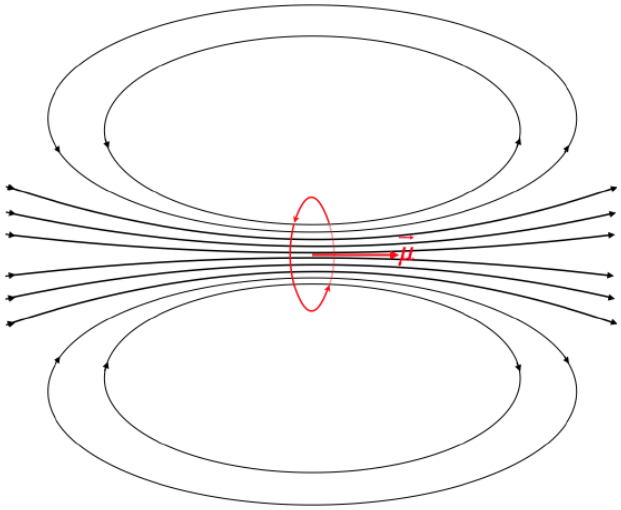
The physical constant that makes the units work out for the force is called the *permeability of free space*:

$$\mu_o = 4\pi \times 10^{-7} \, \tfrac{T \cdot m}{A}$$

Field From a Current in a Circular Wire

If you have a loop of wire instead, the same principles apply as for a straight wire. The diagram below shows a loop wire with current in the counterclockwise direction as viewed from the right side. Using the RHR for currents, we find the that the field lines go through the loop to the right and circle around counterclockwise on the top and clockwise at the bottom. This generates a dipole magnetic field, and like the electric dipole, the field along its axis points in the direction of the dipole moment. Thus, we see that a loop wire is equivalent to a magnetic dipole, with dipole moment, $\vec{\mu}$.

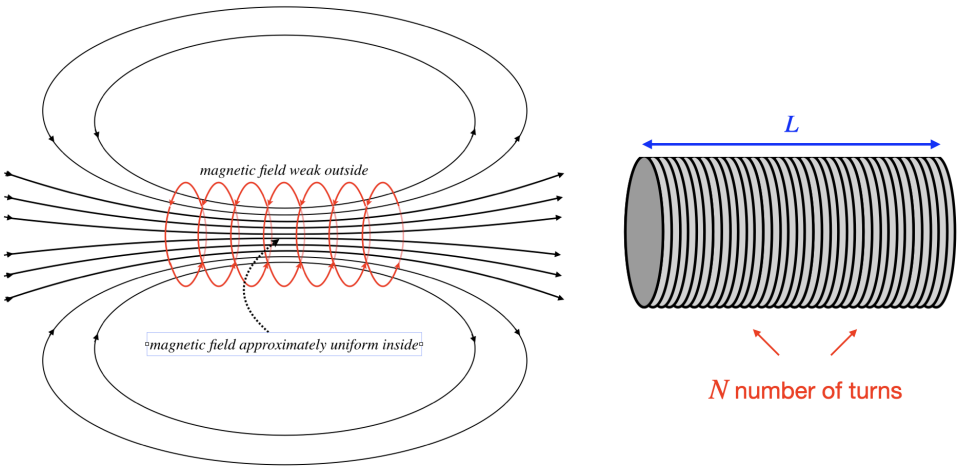
Figure 11.8.3: Magnetic Dipole Field of a Loop



It is possible to stack lots of individual dipoles on top of each other to create a long tube called a *solenoid* as pictured below. By the law of superposition, the magnetic field inside the solenoid keeps adding with each loop, and as the number of loops increases the field inside becomes approximately constant. It can be calculated that the net field inside the solenoid depends on the current, the length of the solenoid, and the number of turns:

$$|\vec{B}| = \mu_o I \frac{N}{L} \tag{11.8.9}$$

Figure 11.8.4: Solenoid Magnetic Field



There are a few particularly interesting aspects of the fields of solenoids:

- The field within the solenoid doesn't change much (it is pretty much *uniform*).
- The field just outside the solenoid (on its side, not the end) is very weak (basically it is *zero*).
- The field looks just like that of a bar magnet, but it can be turned on and off by switching the current on or off.

What is a Magnet?

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 magnetized 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 to be the electric currents produced by the electrons orbiting (i.e. making a current loop) atomic nuclei.

The magnetism of certain materials also depends on spin orientation. Spin endows the electrons with an intrinsic angular momentum. This "intrinsic spin" can only have two possible values (you might have heard that an electron can be "spin up" or

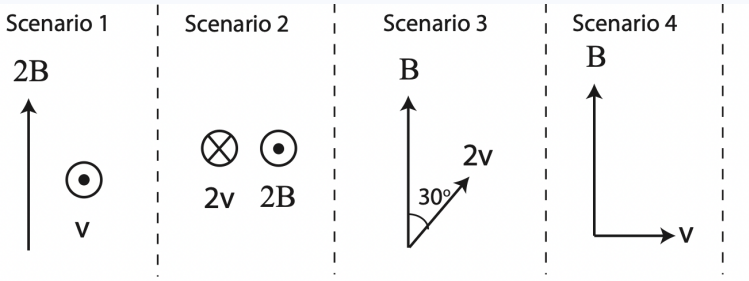
"spin down" in your physics or chemistry classes). Spin is a purely quantum mechanical phenomenon, but for the purposes of thinking about magnetism in our current discussion, consider spin an additional way to produce a loop of current (the smallest one you can imagine!)

To summarize, all our experiments point to the following finding: to get a magnetic field, we need a net motion of charge. 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 field gets weaker. In helium, the spins of the two electrons are paired up-down, so helium would not be very magnetic. As it turns out, 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 which will be covered in a later section.

However, there are many materials whose atoms have an odd number of electrons but who don't exhibit magnetism, how can that be? So far we have discussed the effects of the magnetic field on individual atoms; we need to consider the possibilities that these atoms are also interacting strongly with *each other*. Since a material is made up of $\sim 10^{23}$ atoms, tiny effects, like the ones between atoms, can become large if each atom contributes to it. Highly magnetic materials are generally metals, because metals have many outer electrons that act as if they're "free", and can interact strongly with external fields and with each other.

Example 11.8.1

Shown below are four scenarios depicting directions and magnitudes of magnetic field and velocity of charge. The magnitudes of the charges are the same in all cases, but in Scenario 1 and 3 the charge is positive, and in Scenarios 2 and 4 it is negative.



a) Rank the magnitudes, from smallest to largest, of the magnetic force on the charge in each scenario.

b) Find the direction of the force for each scenario.

Solution

a) Scenario 1: $F = 2vB$

Scenario 2: $F = 0$

Scenario 3: $F = 2vB\sin 30^\circ = vB$

Scenario 4: $F = vB$

Ranking: 2<3=4<1

b) Using the force RHR and flipping answer for negative charges:

Scenario 1: left

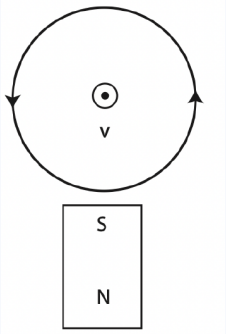
Scenario 2: force is zero

Scenario 3: out of the page

Scenario 4: into the page

Example 11.8.2

In the figure below a negative charge is moving inside a wire with the current flowing counterclockwise, as shown here. A large magnet is also present with the south pole facing the wire. Find the direction of the magnetic force on the charge.

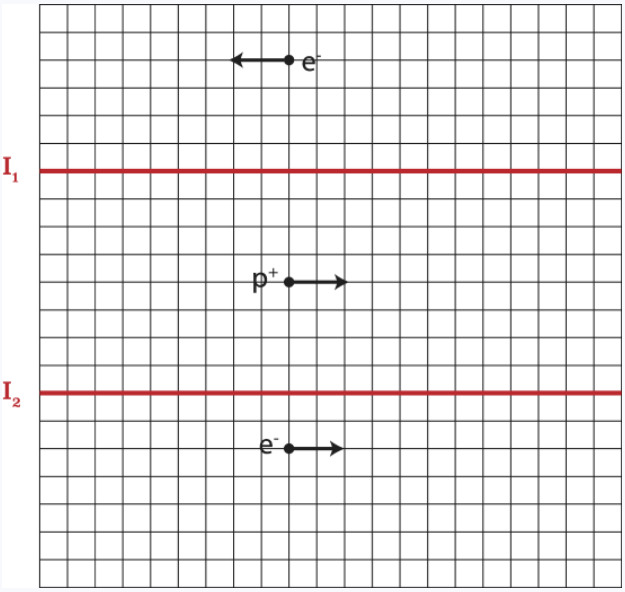


Solution

The RHR for current tells us that the magnetic field due to the wire is point out of the page inside the wire, which is parallel to the motion of the charge, so this field does not contribute to force. Due to the magnet the field points down toward the south pole and away from the north one. Using the RHR for force and flipping answer due to negative charge, we get a magnetic force which points to the left.

Example 11.8.3

Below are two current carrying wires with unknown magnitudes and currents flowing in the opposite directions. Also shown are two electrons and one proton moving in the vicinity of the wires. The arrows represent the direction of motion of each charge. The magnitude and direction of the total magnetic field due to the two wires at the location of the two electrons are equal. The total magnetic field at the location of the proton is out of the page and has a magnitude of $\mu_o \pi$. Each grid square represents a distance of 1m. You may assume that the forces between the charges are negligible.



a) Determine the direction of the current in each wire.

b) Calculate the magnitude of the current in each wire.

c) If the current in wire 2 suddenly turned off, determine the direction of force on the electron above the wire 1.

Solution

a) Since the total magnetic field is out of the page at the location of the proton and the currents are in opposite direction, the top wire's current is to the left and the bottom one is to the right according to the RHR for current.

b) For consistency let's define out of the page as positive direction and into the page as negative. Thus, the total magnetic field at the location of the proton:

$$\vec{B}_{tot} = \vec{B}_1 + \vec{B}_2 = \frac{\mu_o}{(2\pi)(4m)}(I_1 + I_2) = \frac{\mu_o}{\pi}$$

Resulting in:

$$I_1 + I_2 = 8A$$

At the top electron the magnetic field due to wire 1 is into the page and due to wire 2 is out of the page. The magnitude of the total field is:

$$\vec{B}_{tot} = \frac{\mu_o}{2\pi}\left(\frac{I_2}{12} - \frac{I_1}{4}\right)$$

At the bottom electron the magnetic field due to wire 1 is out the page and due to wire 2 is into the page. The magnitude of the total field is:

$$\vec{B}_{tot} = \frac{\mu_o}{2\pi}\left(\frac{I_1}{10} - \frac{I_2}{2}\right)$$

Setting the two equations about equal to each other and solving for I_1 in terms of I_2:

$$I_1 = \frac{35}{21}I_2$$

And plugging into the first equation:

$$\frac{35}{21}I_2 + I_2 = 8A$$

Resulting in:

$$I_2 = 3A; \;\; I_1 = 5A$$

c) Based on result in a) the current in wire 1 is to the left, so the magnetic field is into the page above the wire based on the RHR for wires. Using the RHR for force and flipping result for a negative charge the force is up.

11.9: Magnetic Induction

What Is 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 just a few years later in 1831: *magnetic fields can be used to induce 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 at the time were not impressed as they failed to see the use of it. It turns out that this was the critical step in creating generators and power plants, which make electricity available without flying a kite in a storm or carrying large arrays of batteries. The alternating-current (AC) 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.

Magnetic Flux

Before we tackle 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. Suppose that we want to quantitatively determine the amount of rain that hits the windshield of the car. For simplicity, 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 mainly need to consider these three variables:

- The amount of rain
- The size of the windshield
- The orientation of the windshield relative to the rain

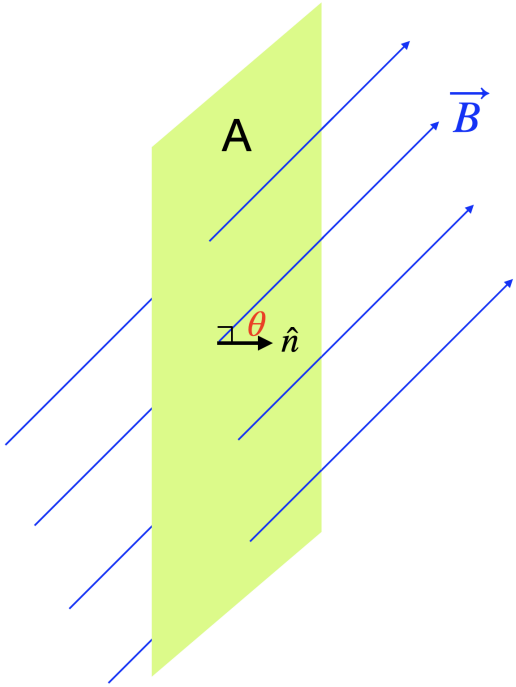
Let’s discuss each in turn. If it is raining hard, there will be a lot more raindrops hitting the windshield than if it is raining lightly. Likewise, If the size of the windshield is large, more raindrops will hit it than if it were small. The orientation between the rain and the windshield will also determine how much rain hits the windshield: if the windshield were arranged vertically, there would be no rain hitting the windshield (in the idealization that the windshield is infinitely thin). Conversely, the most amount of rain will hit if it's arranged perpendicular to the rain, or horizontally (like the sunroof on top of the car). In general, the flux of some quantity through a surface area is a measure of how much of that quantity passes through the area.

This idea of calculating the amount of rain hitting a surface is analogous the concept of *magnetic flux*, Φ. We can consider the flux of the magnetic field passing through an arbitrarily-chosen surface. In our example, the rain falling vertically is our vector field, and the windshield is our arbitrary area. More rain means that the magnitude, $|\vec{B}|$, of the vector field increases. A larger windshield means a greater surface area. The orientation is measured by the angle between the direction of the vector field and the *normal vector*, $\hat{n}$, to the surface area. The normal vector is a unit vector which is perpendicular to the surface through which flux is being calculated. For the particular case of the magnetic field vector $\vec{B}$, we define the magnetic flux Φ through an area A as:

$$\Phi = \vec{B} \cdot A\hat{n} = |\vec{B}||A| \cos\theta \tag{11.9.1}$$

where the angle θ is the angle between the magnetic field vector, $\vec{B}$ and the vector normal, $\hat{n}$ to the surface area A, as depicted in the figure below, for some arbitrary alignment of the magnetic field with a surface represented by a plane.

Figure 11.9.1: Magnetic Flux



From our rain example, you can see that when the rain is falling vertically down, and the windshield surface is horizontal, the normal vector to the area will be vertical. Hence, the rain and the normal vector are parallel to each other, and the angle in Equation 11.9.1 will be $\theta = 0^\circ$. Since $\cos\theta = 1$, this orientation maximizes flux. If the windshield surface is vertical, the vector normal to the surface area will now be horizontal, the angle will be $\theta = 90^\circ$, resulting in $\cos\theta = 0$ and the flux will vanish (no rain hits the vertical windshield). Note that there is 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 by 180°. In other words, the flux as we defined it will change sign, it can be either positive or negative. We will see shortly that changes in flux are what is important physically rather than the actual value of the flux. So any of the choices we make for convention 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 magnetic field, $|\vec{B}|$.
- The size of the surface area, A.
- The angle between the magnetic field vector, $\vec{B}$, and the vector normal, $\vec{n}$, to the area.

Faraday's Law of Induction

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 current in a wire in the presence of the field. For a current to flow, we must have a closed wire loop (a circuit). But unlike in a circuit the loop is not connected to any power source, such as a battery. The *area* we will consider for magnetic flux will be the area enclosed by our wire loop. Note that the wire can be looped in a circular, square, or arbitrarily complicated shape.

Considering the magnetic flux through a wire loop, Faraday asked what happened if you placed a magnet close to the loop and let it sit there. Would a current start to flow in the presence of the magnet? He carried out the experiment and found that there was no current in the loop in this case. However, if you move the magnet away, then for a brief instant a current appears. If you move it back toward the loop, then a current appears again.

What Faraday found is there is an *induced current* (and therefore induced voltage) *only when the magnetic flux changes over time*. We say that the current is *induced* because it's not created by a battery, or some connected voltage source like in a "standard" circuit. The current is induced in the wire by the changing magnetic field. He called the induced voltage the *induced electromotive force*, or induced EMF for short, denoted by $\mathcal{E}$. 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, $\frac{\Delta\Phi}{\Delta t}$.
- If you add loops to the wire coil, each loop will contribute equally to $\mathcal{E}$: if you have N coils, the induced voltage will be N times as strong.

We can now summarize these findings that embody Faraday's Law quantitatively:

$$\mathcal{E} = -N\frac{\Delta\Phi}{\Delta t} \tag{11.9.2}$$

The induced current is the given by:

$$I = \frac{\mathcal{E}}{R} \tag{11.9.3}$$

where R is the resistance in the wire loop. Equation 11.9.2 tells us that we 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. Based on Equation 11.9.1 there are three way that the flux can be changed: changing the strength of the field, $|\vec{B}|$, changing the area, A, (this can be accomplished by part of a loop that is free to move), or by changing θ, a rotation of the field relative to the loop or the loop relative to the field.

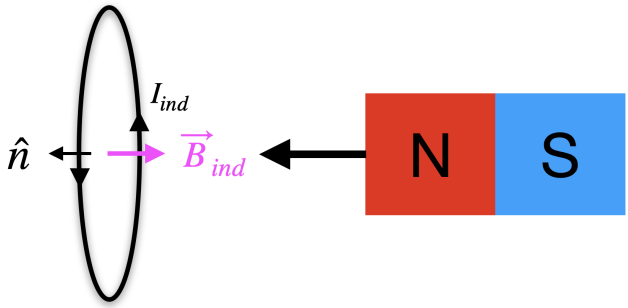
Furthermore, the faster the flux changes, the larger the induced voltage. You can picture this last statement in the following way. If you are inducing current by moving a magnet close to a wire, the current will be larger if you move the magnet quickly than if you move it slowly. The magnitude of the rate of change is proportional to the voltage, the faster the magnetic field changes, the greater the induced current and induced voltage. The negative sign that appears in front of the equation will be explained shortly. Note that Faraday's law focuses only on the effect of a changing magnetic field on a wire. For simplicity, we discussed using a permanent magnet as the source of our field. However, we could also use the magnetic field produced by current in another wire. In fact, this is how Faraday studied induced current and induced voltages.

Lenz's Law: The Direction of Induced Current

Equation 11.9.2 for the EMF, $\mathcal{E}$, derived from Faraday's Law, also has a sign. What is its significance? The sign gives the *direction* of the induced current in the loop. So far we have not discussed how we are to choose between the two possibilities for the current's direction. Experimentally, if we change a magnetic flux to induce a current, the current will flow to produce a *new* flux that *opposes* our change. This is known as *Lenz's Law*. The best way to understand Lenz's Law is by looking at a few examples.

Consider a circular loop of wire in the figure below. Initially, there is no magnetic field in the region of the wire. At $t = 0$ we create a changing magnetic field by moving a bar magnet toward the loop as shown. We make the magnitude of the field increase *linearly* with time. From Faraday's Law, we know that there will be an induced voltage, because the flux through the loop is increasing as the field strength increases.

Figure 11.9.2: Changing Flux Example



The area enclosed by the circuit is constant. The angle between the normal to the area of the circle loop and the field is constant, $\theta = 0$ and not changing. The magnitude of the magnetic field, $|\vec{B}|$, field increases with time. Therefore, the changing magnetic field is the only contribution to the change in flux:

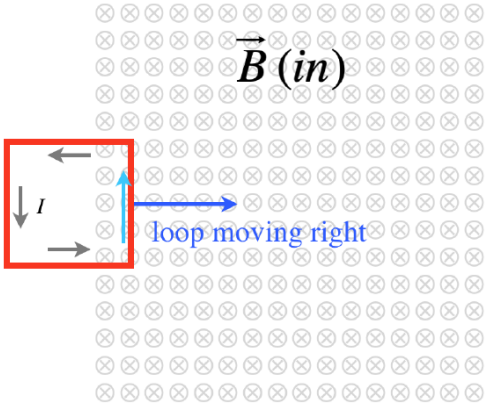
$$\frac{\Delta\Phi}{\Delta t} = -A\frac{\Delta B}{\Delta t} \tag{11.9.4}$$

We can turn the derivative in Equation 11.9.1 into just changes, "Δs", since we are assuming that the field is changing linearly with time. Since the magnitude increases linearly with time, the rate of change the magnetic field magnitude versus time is a constant: therefore the induced voltage will be constant.

The question we want tot answer is: will the current flow clockwise or counterclockwise? Now we use Lenz's Law. To consider flux, let's choose the normal to be parallel to the magnetic field. In this example, since the magnetic field points from the north side of the magnet, this means that the normal vector will point to the left (as shown in the figure). This means the flux is positive and it increases with time. Based on Lenz's Law the induced current should *oppose* this. Thus, it will produce a *negative* magnetic flux through the loop that tries to cancel the increase in the positive flux. Since a positive flux means the field points to the left, then the magnetic field produced by the induced current, which we call the *induced field*, $\vec{B}_{ind}$, will point to the right. Now can use the current right-hand-rule to established the direction of current based on the magnetic field, $\vec{B}_{ind}$, it produces. When you align your thumb along the tangent to the loop, and curl your fingers towards the induced magnetic field, your thumb will indicate the direction of the induced current. Since the induced field points to the right, we see that the induced current for this case is flowing in a counterclockwise direction as viewed from the right.

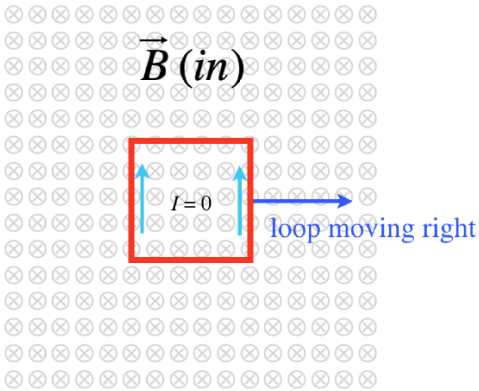
Below is another example of a conducting loop moving to the right into a uniform field which points into the page. Initially, there is no field through the loop. As it enters the field there is a flux into the page. To oppose this change the induced field will be out of the page. According to the right-hand-rule, this implies that the induced current will be counterclockwise as shown.

Figure 11.9.3: Conducting Loop Enters Uniform Magnetic Field



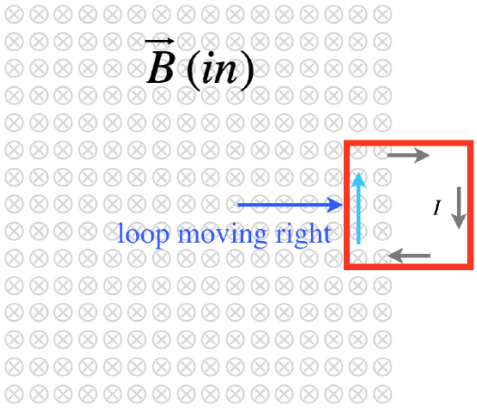
Next, as the loop continues to move to the right within the field, there is no longer a change in flux, this results in zero induced field and thus no current will flow.

Figure 11.9.4: Conducting Loop Stays Within the Uniform Magnetic Field



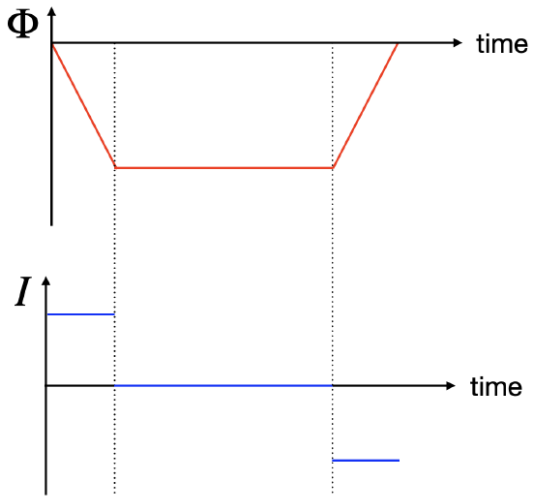
Finally, as the loop starts to exit the magnetic field, the flux that points into the page starts to decrease. To oppose this change, the induced flux and thus field will point into the page. With the help of the right-hand-rule we find that this must result in a clockwise induced current.

Figure 11.9.5: Conducting Loop Exits the Uniform Magnetic Field



We can explore directly how the changing flux is related to current by making a plot of flux as a function of time, as depicted in the figure below. Flux will always be into the page for this scenario since the magnetic field points into the page. If we will define the normal vector to point out of the page, the flux will be negative since the angle between flux and the normal vector is 180^o in this example. We also assume that flux changes linearly as the loop moves to the right at a constant speed. The flux starts at zero and its magnitude increases as the loop moves into the field. Since the flux is negative, this means that it becomes more negative. The constant flux region in the plot represent the part of the motion when the loop is fully immersed in the field, thus, the amount of flux does not change as it moves through it. As it starts exiting the field, the magnitude of flux starts to decrease or become less negative until it reaches zero when it fully exists the field.

Figure 11.9.6: Flux and Current as a Function of Time



The plot also shows the induced current as a function of time. Equation 11.9.2 tells us that the induced current is the negative of the derivative of flux as a function of time. Initially, since the slope of the flux is constant and negative, this results in a constant positive current. This means that by defining flux into the page as negative, the change in flux will be also negative, so the induced field that opposes this change has to be positive. This leads to a counterclockwise induced current which by convention is then defined as positive. In the region where the flux is constant, the slope is zero, resulting in zero induced current. When the flux starts to decrease (become less negative), the slope is positive, resulting in a negative (clockwise) current.

Applications

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 phenomenon has found many familiar applications in the modern world:

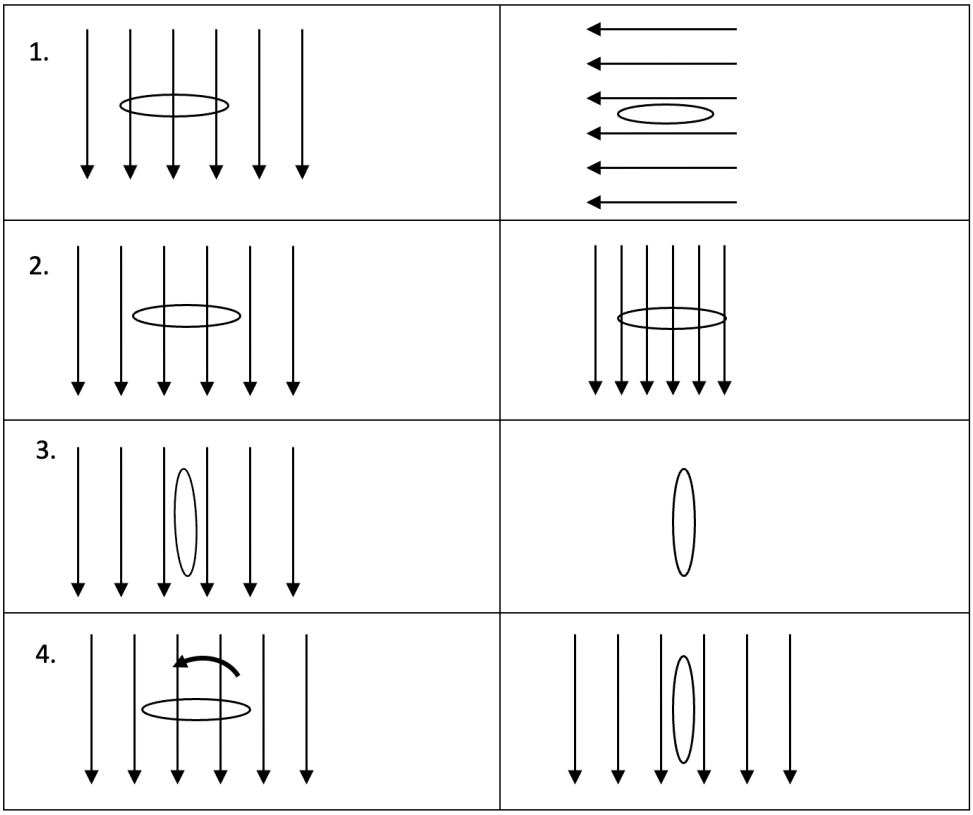
- **Seismograph:** One way to exploit Faraday's Law is to attach a magnet to anything that moves and place it near a loop of wire; any movement or oscillation in the object can be detected as an induced current in the wire loop. In this way we can translate physical movements and oscillations into electrical 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, as the vibrations produced by an earthquake pass through a seismograph, a magnet's vibrations produce a current that can be amplified to drive a plotting pen. This is how the 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, it vibrates, creating a changing magnetic flux through the coil of wire surrounding the permanent magnet. The coil "picks up" the vibrations that generate an induced current and sends the signal 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 (like in a hydroelectric generator), expanding steam (as in coal, oil, and nuclear power plants), or wind (as in wind turbine generators). In all cases, the principle is the same, the mechanical energy is used to move a conducting wire coil inside a magnetic field (usually by rotating the wire). In this case, the area of the coil is the constant, the magnitude of the field is constant, so the angle term in the equation for Faraday's law is responsible for the changing flux. This is caused by the change in the relative orientation between the magnetic field and the normal to the area of the coil. 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 $\cos \omega t$. Using calculus, the time rate of change of the flux will then be proportional to $\omega \sin \omega t$. This means the induced current will oscillate *sinusoidally*. In other words, the current in the coil alternates in direction, flowing in one direction for half the cycle and flowing the other direction for the other half. This kind of generator is referred to as 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:** Electric motors work in basically the reverse principle that operates electric generators: an alternating electric current causes an electromagnetic cylinder to periodically switch poles, which interacts with the field of an inlaid magnet to turn it. Some motors use electromagnets for both components, but the principle is the same. The stationary magnetic piece is called the *stator* and the magnetic piece that rotates is called that *rotor*
- **Hybrid Cars**: Regenerative breaking discussed above.

Example 11.9.1

Below are 4 cases: the left panels show the initial and the right panels show the final configurations. The arrows indicate the direction and magnitude of the external magnetic field.



Fill in the table below for each case

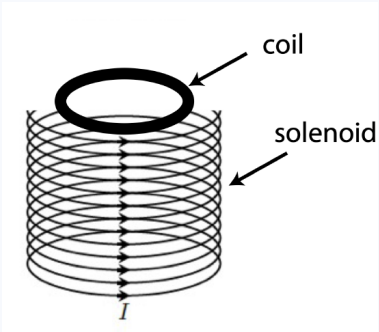
Case	Change in Flux (Yes or No). If yes, what changed (A, B, or θ)	Direction of induced Field	Direction of induced Current	Brief Explanation
1				
2				
3				
4				

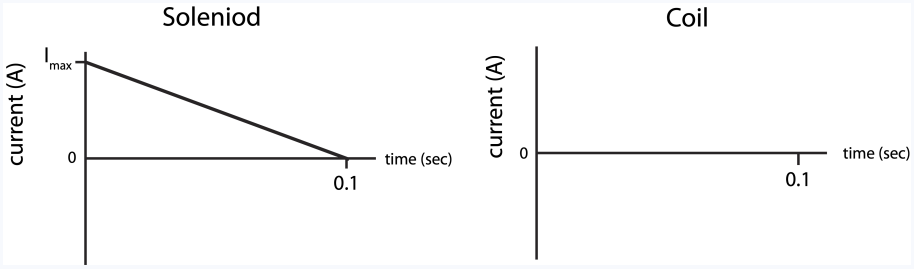
Solution

Case	Change in Flux (Yes or No). If yes, what changed (A, B or θ)	Direction of induced Field	Direction of induced Current	Brief Explanation
1	*Yes: angle*	*Down*	*Clockwise from above*	*B field down goes to zero, so the induced field is down. Using RHR results in CW current from above.*
2	*Yes: B field*	*Up*	*Counterclockwise from above*	*Density of field down increases, to oppose that induced field is up. Using RH results in CCW current from above.*
3	*No*	*N/A*	*N/A*	*Flux is not changing, so there is no induced current*
4	*Yes: angle*	*To the right*	*Clockwise from the left*	*B field down goes to zero, so induced field is to the right as the loop rotates. Using RHR results in CW current viewed from the left.*

Example 11.9.2

Consider a magnetic field is created by a large solenoid magnet. The solenoid is 1.5 meters long, has 5000 turns, a resistance of 4Ω, and a one-meter radius. A coil located inside the solenoid has a single loop (as depicted below), a resistance of 0.6Ω, and a 0.4 m radius. The solenoid initially has a current which produces a 0.2 T magnetic field. The solenoid's current is then reduced linearly to zero in 0.1 seconds as shown in the left plot below.





a) Calculate I_{max} which is marked on the plot below.

b) Make a graph of the current in the coil on the right plot below. Make sure to indicate numerical values and explain your choice of sign for the current.

Solution

a) The magnetic field for a solenoid is $B = \frac{\mu_o I N}{L}$. The maximum current with be when the magnetic field is at 0.2 T, since the current is reduced with time:

$$I_{max} = \frac{BL}{\mu_o N} = \frac{0.2T \times 1.5m}{4 \times 10^{-7} \frac{N}{A^2} \times 5000} = 47.7\ A$$

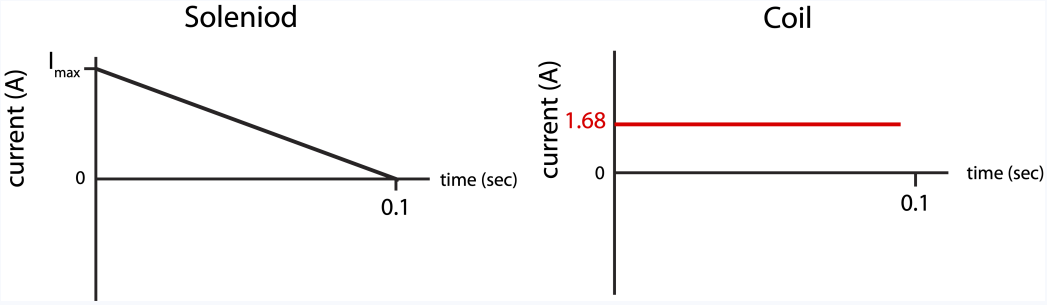
b) The magnetic field is upward using the current RHR, which is reduced to zero. Thus, the induced magnetic field will also be upward to oppose the change of flux decreasing in the upward direction. Using the same RHR again, this results in an induced current which is counterclockwise as viewed from top. In other words, the direction of the current in the coil is the same as the direction of the current in the solenoid.

Now that we have established the direction of current, we just need to worry about the magnitude of the induced emf and current. The magnitude of induced emf for the wire which only contains one loop is:

$$|\mathcal{E}| = \frac{d\Phi}{dt} = A\frac{\Delta B}{\Delta t} = \pi r^2 \frac{dB}{dt} = (0.4^2\pi)m^2 \frac{0.2T}{0.1s} = 1.68A$$

Solving for induced current:

$$I_{coil} = \frac{|\mathcal{E}|}{R} = \frac{1.0V}{0.6\Omega} = 1.68A$$



Index

D

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: UCD: Physics 7C - General Physics

Webpages: 138

All licenses found:

- Undeclared: 100% (138 pages)

By Page

- UCD: Physics 7C - General Physics - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 8: Waves (old version) - *Undeclared*
    - 8.1: Introduction to Waves - *Undeclared*
      - Front Matter - *Undeclared*
        - TitlePage - *Undeclared*
        - InfoPage - *Undeclared*
      - 1. What is a Wave? - *Undeclared*
      - 2. Wave Properties and Characteristics - *Undeclared*
      - 3. Harmonic Waves - *Undeclared*
      - 4. Graphical Representations of Waves - *Undeclared*
      - 5. Other Types of Waves - *Undeclared*
      - 6. Summary - *Undeclared*
      - Back Matter - *Undeclared*
        - Index - *Undeclared*
    - 8.2: Wave Interference and Superposition - *Undeclared*
      - Front Matter - *Undeclared*
        - TitlePage - *Undeclared*
        - InfoPage - *Undeclared*
      - 1. Superposition Overview and Basics - *Undeclared*
      - 2. Superposition of Harmonic Waves - *Undeclared*
      - 3. Standing Waves - *Undeclared*
      - 4. Beats - *Undeclared*
      - 5. Two-Slit Interference - *Undeclared*
      - 6. Summary - *Undeclared*
      - Back Matter - *Undeclared*
        - Index - *Undeclared*
    - 8.3: Quantum Mechanics - *Undeclared*
      - 1. Introduction - *Undeclared*
      - 2. Quantized Energies - *Undeclared*
      - 3. What are Matter Waves? - *Undeclared*
      - 4. Summary - *Undeclared*
  - 8: Waves - *Undeclared*
    - 8.1: Introduction to Waves - *Undeclared*
    - 8.2: Wave Representation - *Undeclared*
    - 8.3: Multi-Dimensional Waves - *Undeclared*
    - 8.4: Doppler Effect - *Undeclared*
    - 8.5: Superposition and Interference - *Undeclared*
    - 8.6: Beats - *Undeclared*
    - 8.7: Double-Slit Interference - *Undeclared*
    - 8.8: Standing Waves - *Undeclared*
  - 9: Optics (old version) - *Undeclared*
    - 9.1: Rays and Wavefronts - *Undeclared*
    - 9.2: Optics and Images - *Undeclared*
      - Front Matter - *Undeclared*
        - TitlePage - *Undeclared*
        - InfoPage - *Undeclared*
      - 9.2.1: Overview of Geometric Optics - *Undeclared*
      - 9.2.2: Reflection - *Undeclared*
      - 9.2.3: Refraction - *Undeclared*
      - 9.2.4: Total Internal Reflection - *Undeclared*
      - 9.2.5: Images Again - *Undeclared*
      - 9.2.6: Review - *Undeclared*
      - Back Matter - *Undeclared*
        - Index - *Undeclared*
    - 9.3: Lenses - *Undeclared*
      - Front Matter - *Undeclared*
        - TitlePage - *Undeclared*
        - InfoPage - *Undeclared*
      - 9.3.1: Overview of Lenses - *Undeclared*
      - 9.3.2: Lenses and Ray Tracing - *Undeclared*
      - 9.3.3: The Thin Lens Equation - *Undeclared*
      - 9.3.4: Multiple Lenses - *Undeclared*
      - 9.3.5: Applications - *Undeclared*
      - Back Matter - *Undeclared*
        - Index - *Undeclared*