UCD: PHYSICS 9HC – INTRODUCTION TO WAVES, PHYSICAL OPTICS, AND QUANTUM THEORY

Tom Weideman
University of California, Davis

LibreTexts™

University of California, Davis

UCD: Physics 9HC – Introduction to Waves, Physical Optics, and Quantum Theory

Tom Weideman

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

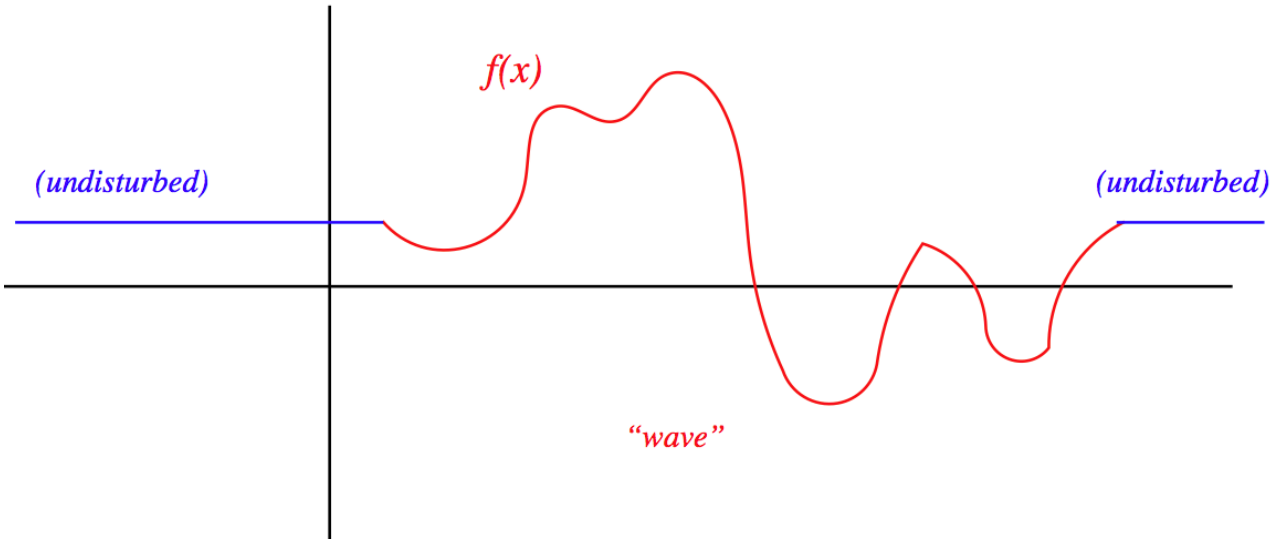
1: Waves

1.1: Wave Mathematics

Definition of a Wave

A *wave* is a disturbance that propagates through space at a constant speed. With one notable exception that we will encounter later, this "disturbance" consists of a fluctuation in the ambient condition of a medium. Waves in one dimension maintain a consistent *waveform* as they propagate (later we will see why this is not so for waves in two and three dimensions). Let's see how we can model this mathematically. We'll start with a localized disturbance frozen in time (think of it as a snapshot) that we describe with a function $f(x)$:

Figure 1.1.1 – Snapshot of a Wave



This is the waveform, but to be a wave, it needs to be propagating along the x-axis, which would make it a function of both x and t. To turn it into such a function, we first have to think about how a function can be shifted along the x-axis. This is accomplished by replacing x in the argument of the function with the sum or difference of x and the value of the shift. If one wishes to shift the function $f(x)$ in the $+x$ direction by a distance a, then the proper change is to the function $f(x-a)$. Note that *subtracting* a in the argument shifts the function in the *positive* x direction, and adding the constant shifts it in the negative x direction. We insist that the wave moves at a constant speed, so we want the wave form to shift by the same distance every time the same time interval passes. We therefore have that the general form of a wave function is:

$$f(x,t) = f(x \pm vt) \tag{1.1.1}$$

This represents a waveform $f(x)$ propagating in the $\mp x$ direction with a speed v.

There are countless functions of x and t that we can come up with, but not all can be written in the form described above. When faced with an arbitrary function of x and t, it can be challenging to determine whether the function represents a wave.

Example 1.1.1

Determine which (if any) of the functions below represent a traveling wave. For those that do, determine the direction of their propagation, and their speed. In every case the constants α and β are positive numbers.

a. $f(x,t) = (1-\alpha x+\beta t)^3 + (2+\alpha x-\beta t)^5$
b. $f(x,t) = \sin\left[(\alpha x)^2 - (\beta t)^2\right]$
c. $f(x,t) = e^{-\alpha x} e^{-\beta t}$

Solution

The idea here is to do whatever algebra that is necessary to get the function into the form $f(x \pm vt)$ ...

a. If we factor $-\alpha^3$ out of the first term, and α^5 out of the second term, we have:

$$f(x,t) = -\alpha^3\left(-\frac{1}{\alpha} + \left(x - \frac{\beta}{\alpha}t\right)\right)^3 + \alpha^5\left(\frac{2}{\alpha} + \left(x - \frac{\beta}{\alpha}t\right)\right)^5$$

We can see that this is purely a function of $\left(x - \frac{\beta}{\alpha}t\right)$. In such problems, it might help to substitute z for $(x \pm vt)$ and show that there are no x's or t's left over. The resulting function $f(z)$ is in fact the waveform. For this case:

$$f(z) = (1-\alpha z)^3 + (2+\alpha z)^5$$

This is therefore a traveling wave moving in the $+x$ direction (because of the opposite signs of x and t), and the speed must be $\frac{\beta}{\alpha}$.

b. At first glance this might appear to be the function of a traveling wave, but we can show that in fact it cannot be written in the correct form. Writing the difference of two squares as a product gives:

$$f(x,t) = \sin[(\alpha x + \beta t)(\alpha x - \beta t)]$$

Each of the factors has the right form (once a factor of α is divided out), but if we substitute z for one of them, we cannot similarly eliminate the second factor. Put another way, one of the factors indicates the wave is moving in the $+x$ direction, while the other indicates it is moving the opposite way. It can't be doing both, so this is not the equation of a traveling wave.

c. Combining the exponentials gives:

$$f(x,t) = e^{-\alpha x - \beta t} = e^{-\alpha\left(x + \frac{\beta}{\alpha}t\right)} \quad \Rightarrow \quad f(z) = e^{-\alpha z}$$

Clearly this represents a wave propagating in the $-x$ direction with a speed of $\frac{\beta}{\alpha}$.

The Wave Equation

It seems like there has to be an easier way to determine if a function of x and t represents a wave. It turns out that there is! To see this, let's start with the basic definition above. If we define $g_\pm(x,t) \equiv x \pm vt$, then we can write the wave function as $f(g_\pm)$. Now we can write derivatives of the function with respect to x and t in terms of derivatives with respect to $g_\pm$ using the chain rule. Note that these are functions of more than one variable, so we need to use *partial* derivatives. These work precisely like ordinary derivatives, except that when the derivative is taken with respect to one variable, all the other variables are treated as constants.

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{\partial f}{\partial g_\pm}\frac{\partial g_\pm}{\partial x} = \frac{\partial f}{\partial g_\pm} \quad &\Rightarrow \quad \frac{\partial^2 f}{\partial x^2} &= \frac{\partial^2 f}{\partial g_\pm^2} \\ \frac{\partial f}{\partial t} &= \frac{\partial f}{\partial g_\pm}\frac{\partial g_\pm}{\partial t} = \pm v\frac{\partial f}{\partial g_\pm} \quad &\Rightarrow \quad \frac{\partial^2 f}{\partial t^2} &= v^2\frac{\partial^2 f}{\partial g_\pm^2} \end{aligned} \tag{1.1.2}$$

Putting these together gives us a relation between second derivatives known as the *wave equation*:

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{1.1.3}$$

This second order partial differential equation holds if and only if the function behaves like a traveling wave (or a linear combination of traveling waves) with speed v.

Example 1.1.2

For the functions in the previous example, demonstrate whether they do or do not satisfy the wave equation with the proper wave speed.

Solution

We found that the formulas for cases (a) and (c) represent waves, so we plug those into the wave equation:

(a)

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{\partial^2}{\partial x^2}\left[(1-\alpha x+\beta t)^3 + (2+\alpha x-\beta t)^5\right] = \frac{\partial}{\partial x}\left[-3\alpha(1-\alpha x+\beta t)^2 + 5\alpha(2+\alpha x-\beta t)^4\right] = 6\alpha^2(1-\alpha x+\beta t) + 20\alpha^2(2+\alpha x-\beta t)^3$$

$$\frac{\partial^2}{\partial t^2} f(x,t) = \frac{\partial^2}{\partial t^2}\left[(1-\alpha x+\beta t)^3 + (2+\alpha x-\beta t)^5\right] = \frac{\partial}{\partial t}\left[3\beta(1-\alpha x+\beta t)^2 - 5\beta(2+\alpha x-\beta t)^4\right] = 6\beta^2(1-\alpha x+\beta t) + 20\beta^2(2+\alpha x-\beta t)^3$$

From direct comparison, it is clear that these two terms are proportional, which means they satisfy the wave equation:

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{\alpha^2}{\beta^2}\frac{\partial^2}{\partial t^2} f(x,t)$$

The constant of proportionality for the wave equation is $\frac{1}{v^2}$, so this confirms that $v = \frac{\beta}{\alpha}$.

(b)

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{\partial^2}{\partial x^2} \sin\left[(\alpha x)^2 - (\beta t)^2\right] = \frac{\partial}{\partial x}\left\{2\alpha^2 x \cos\left[(\alpha x)^2 - (\beta t)^2\right]\right\} = 2\alpha^2 \cos\left[(\alpha x)^2 - (\beta t)^2\right] - 4\alpha^4 x^2 \sin\left[(\alpha x)^2 - (\beta t)^2\right]$$

$$\frac{\partial^2}{\partial t^2} f(x,t) = \frac{\partial^2}{\partial t^2} \sin\left[(\alpha x)^2 - (\beta t)^2\right] = \frac{\partial}{\partial t}\left\{-2\beta^2 t \cos\left[(\alpha x)^2 - (\beta t)^2\right]\right\} = -2\beta^2 \cos\left[(\alpha x)^2 - (\beta t)^2\right] - 4\beta^4 t^2 \sin\left[(\alpha x)^2 - (\beta t)^2\right]$$

These two terms are clearly not proportional, so this function does not satisfy the wave equation.

(c)

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{\partial^2}{\partial x^2}\left[e^{-\alpha x} e^{-\beta t}\right] = \frac{\partial}{\partial x}\left[-\alpha e^{-\alpha x} e^{-\beta t}\right] = \alpha^2 e^{-\alpha x} e^{-\beta t}$$

$$\frac{\partial^2}{\partial t^2} f(x,t) = \frac{\partial^2}{\partial t^2}\left[e^{-\alpha x} e^{-\beta t}\right] = \frac{\partial}{\partial t}\left[-\beta e^{-\alpha x} e^{-\beta t}\right] = \beta^2 e^{-\alpha x} e^{-\beta t}$$

These two terms are proportional, and the constant of proportionality gives the correct velocity once again.

Waves in Two and Three Dimensions

Consider a two-dimensional wave, such as a ripple radiating outward from a pebble dropped into a still lake. If the distance of the wave front from the source is r, following the "wave form remains unchanged" prescription, the functional form of the traveling wave is $f(r,t) = f(r - vt)$. Alternatively, we can extend the wave equation to two or three dimensions as follows:

$$\begin{aligned} &\text{two dimensions:} && \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \\ &\text{three dimensions:} && \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \end{aligned} \tag{1.1.4}$$

In the one-dimensional case, we showed above that the wave form that remains unchanged as it propagates is described by a function that satisfies the wave equation. It turns out that in the two-dimensional case, this is no longer true. We can show this by repeating the procedure outlined in Equation 1.1.2. A waveform that remains unchanged as it spreads radially outward would have the form $f(r,t) = f(r - vt)$ (we are considering an outgoing waveform, which accounts for the minus sign). Defining the function $g(r,t) \equiv r - vt$, we have for the derivative of the wave function with respect to x:

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial x}\right) = \left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial r}\right)\left(\frac{\partial r}{\partial x}\right) \tag{1.1.5}$$

Note that the variable r depends upon both x and y, specifically:

$$r = \sqrt{x^2 + y^2} = \left(x^2 + y^2\right)^{\frac{1}{2}} \quad \Rightarrow \quad \frac{\partial r}{\partial x} = \frac{1}{2}\left(x^2 + y^2\right)^{-\frac{1}{2}}(2x) = \frac{x}{r}, \quad \text{and} \quad \frac{\partial r}{\partial y} = \frac{y}{r} \tag{1.1.6}$$

Plugging this and $\frac{\partial g}{\partial r} = 1$ in above gives:

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{\partial f}{\partial g}\left(\frac{x}{r}\right) \quad &\Rightarrow \quad \frac{\partial^2 f}{\partial x^2} &= \frac{\partial^2 f}{\partial g^2}\left(\frac{x^2}{r^2}\right) + \frac{\partial f}{\partial g}\left(\frac{1}{r}\right)\left(1 - \frac{x^2}{r^2}\right) \\ \frac{\partial f}{\partial y} &= \frac{\partial f}{\partial g}\left(\frac{y}{r}\right) \quad &\Rightarrow \quad \frac{\partial^2 f}{\partial y^2} &= \frac{\partial^2 f}{\partial g^2}\left(\frac{y^2}{r^2}\right) + \frac{\partial f}{\partial g}\left(\frac{1}{r}\right)\left(1 - \frac{y^2}{r^2}\right) \end{aligned} \tag{1.1.7}$$

The dependence on t is the same as in the one-dimensional case:

$$\frac{\partial f}{\partial g} = -\frac{1}{v}\frac{\partial f}{\partial t} \quad \Rightarrow \quad \frac{\partial^2 f}{\partial g^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{1.1.8}$$

Plugging this into the previous equations, and adding them together gives:

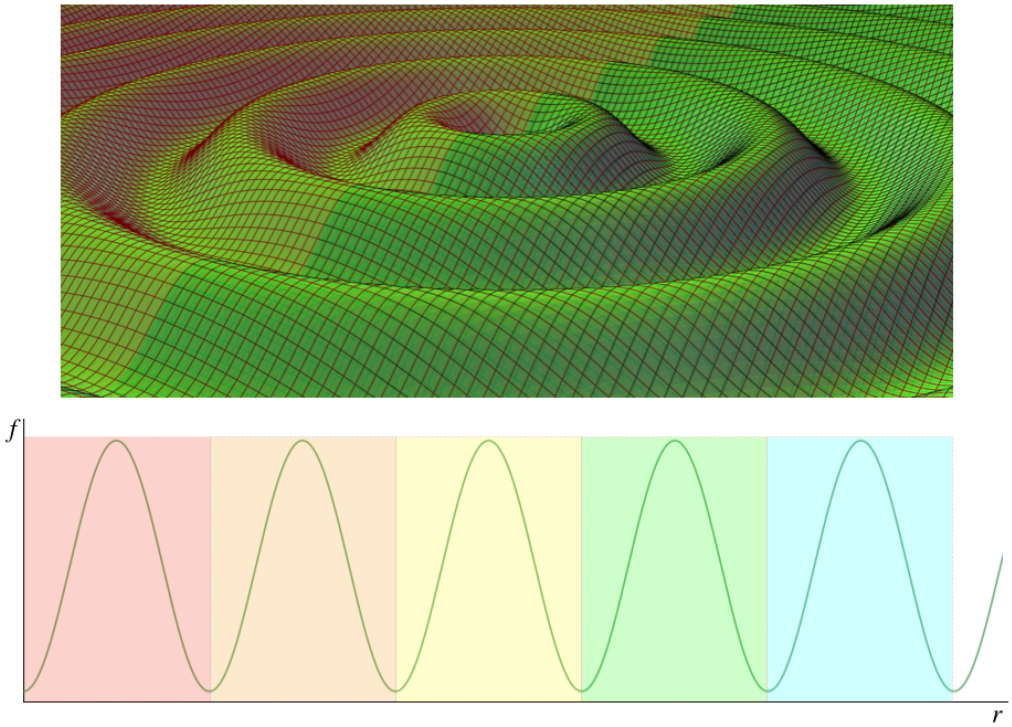
$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}\left(\frac{x^2}{r^2} + \frac{y^2}{r^2}\right) - \frac{1}{v}\frac{\partial f}{\partial t}\left(\frac{1}{r}\right)\left(2 - \frac{x^2}{r^2} - \frac{y^2}{r^2}\right) \tag{1.1.9}$$

Noting that $x^2 + y^2 = r^2$, we finally get:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} - \frac{1}{v}\frac{\partial f}{\partial t}\left(\frac{1}{r}\right) \tag{1.1.10}$$

The second term on the right hand side of the equation clearly makes this differential equation look different from the one-dimensional wave equation. So the question is, which is the correct way of describing a two-dimensional wave? Does it maintain its wave form as it propagates outward, or does it satisfy our previous wave equation extended to two dimensions? The figures below display the two possibilities we are talking about.

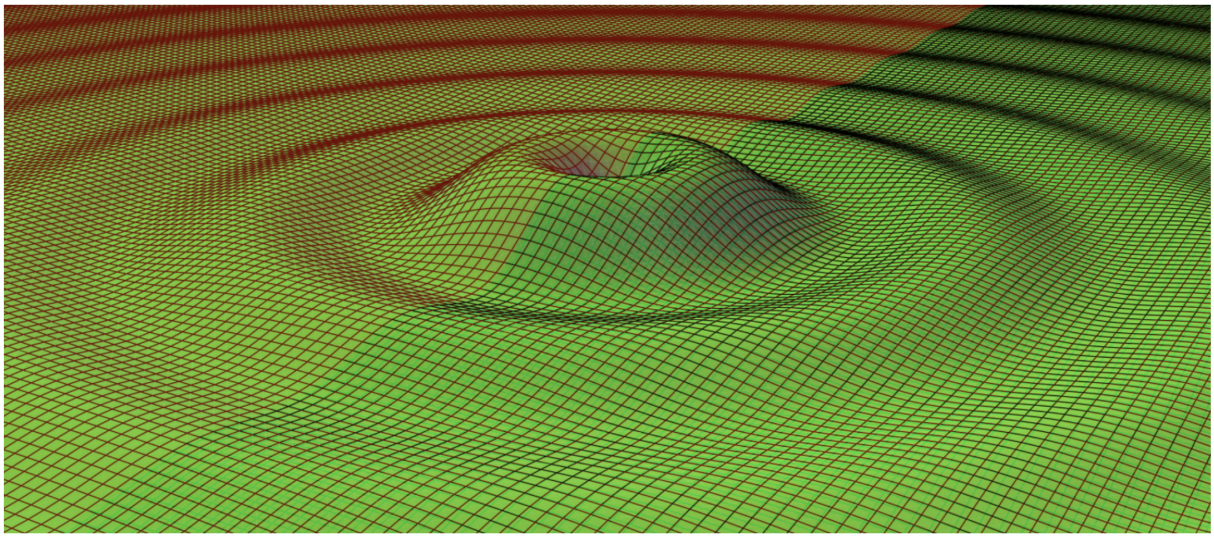
Figure 1.1.2 – Two-Dimensional Circularly-Radiating Function with Unchanging Waveform

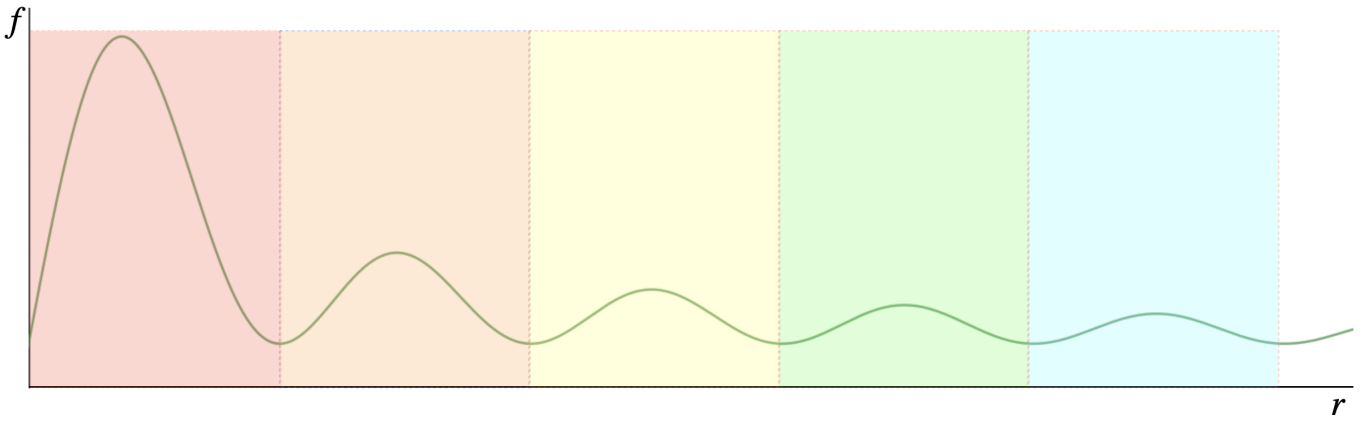


wave form remains the same as wave moves away from center: $f(r,t) = f(r - vt)$

The graph shows a cross-sectional snapshot of the wave – the waveform repeats as a function of r.

Figure 1.1.3 – Two-Dimensional Circularly-Radiating Function Satisfying 2-D Wave Equation





wave form diminishes in height as wave moves away from center: $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}$

The graph shows a cross-sectional snapshot of the wave – the waveform does not repeat as a function of r.

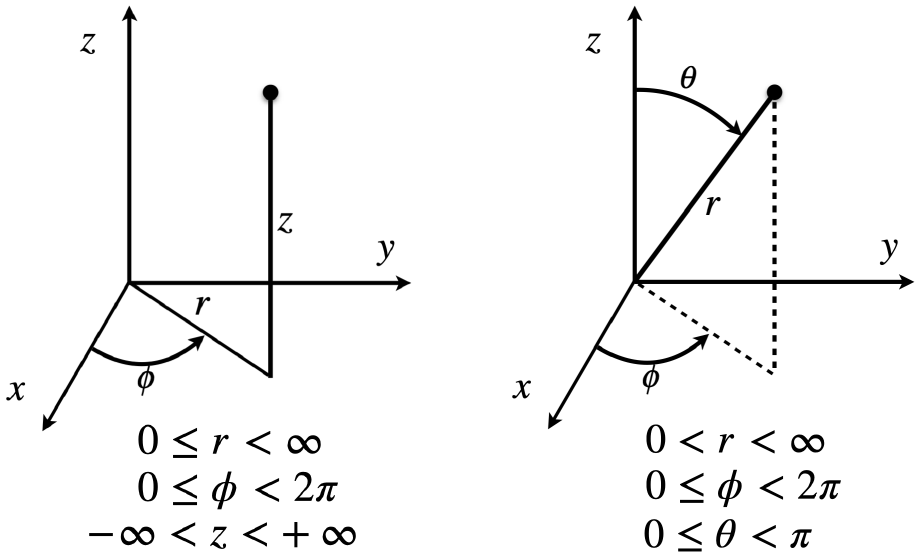
We can't answer this question purely mathematically – we have to observe what actually happens in nature. As we will see in later section, conservation of energy will require that it is in fact the extended two-dimensional wave equation that gives the correct answer – two and three-dimensional waveforms *do not remain fixed* as they radiate outward. The maximal wave displacements diminish with distance from the source, as shown in Figure 1.1.3. Of course, two-dimensional waves don't have to move purely radially outward from a central source, but solutions to the wave equation become extremely complicated in such cases, so we will never address them.

The extension to the three-dimensional wave equation should be obvious:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{1.1.11}$$

As with the two-dimensional case, radially-moving waves in three dimensions have diminishing maximal wave displacements, for the same reason (energy conservation). It turns out that a change of coordinate systems makes it much easier to discuss waves that travel purely radially. In two dimensions, the useful coordinate system is cylindrical coordinates, and in three dimensions it is spherical coordinates:

Figure 1.1.4 – Cylindrical and Spherical Coordinates



$$0 \le r < \infty \qquad 0 < r < \infty$$
$$0 \le \phi < 2\pi \qquad 0 \le \phi < 2\pi$$
$$-\infty < z < +\infty \qquad 0 \le \theta < \pi$$

Note that the two r variables have different definitions in these coordinate systems, so one must take care to keep track of the context when these variables are in play. Specifically, we can write each of these r's in terms of the Cartesian coordinates as follows:

$$\text{cylindrical:}\quad r = \sqrt{x^2+y^2} \qquad \text{spherical:}\quad r = \sqrt{x^2+y^2+z^2} \tag{1.1.12}$$

A little bit of trigonometry yields a translation between the angles and the Cartesian coordinates as well:

$$\phi = \cos^{-1}\frac{x}{\sqrt{x^2+y^2}} \qquad \theta = \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}} \tag{1.1.13}$$

While we certainly have no reason to produce it here, these translations between coordinate systems allows us to derive the wave equation in each of these other coordinate systems. The result is a huge mess, but as we have only considered radial waves here, we can simplify this mess. A radially-moving wave does not vary in the angular directions (nor in the 2-dimensional case, in the z-direction – it has *only* radial dependence). Therefore the only derivatives for the wave equation under these special circumstances that survive are with respect to the variable r. The resulting radial wave equations are:

$$\text{cylindrical:}\quad \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)f(r,t) = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}f(r,t) \qquad \text{spherical:}\quad \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)f(r,t) = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}f(r,t) \tag{1.1.14}$$

It turns out that although these wave equations don't result in a sustained wave form like the 1-dimensional case, we can nevertheless write the radially-moving wave solutions in terms of a function that looks like $f(r \pm vt)$. For the spherical case, this expression is exact, but for the cylindrical case it is only an approximation that gets better as r gets larger. Doing this shows explicitly the property of reduced heights of wave peaks, since the wave form is in the numerator, and the denominator grows (making the total wave displacement smaller) with increased distance from the center:

$$\text{cylindrical:}\quad f(r,t) \approx \frac{f(r \pm vt)}{\sqrt{r}} \qquad \text{spherical:}\quad f(r,t) = \frac{f(r \pm vt)}{r} \tag{1.1.15}$$

An Important Feature of the Wave Equation

One feature of the wave equation that we will use over and over is the fact that it is *linear*. What this means is that if two different functions satisfy the same wave equation, then so does their sum. Or more generally, so does a *linear combination* of those functions. This is true in any number of dimensions, but is quite obvious in one dimension:

$$\left.\begin{aligned} a\left(\frac{\partial^2 f_1}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 f_1}{\partial t^2}\right) \\ b\left(\frac{\partial^2 f_2}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 f_2}{\partial t^2}\right) \end{aligned}\right\} \quad a\frac{\partial^2 f_1}{\partial x^2} + b\frac{\partial^2 f_2}{\partial x^2} = a\frac{1}{v^2}\frac{\partial^2 f_1}{\partial t^2} + b\frac{1}{v^2}\frac{\partial^2 f_2}{\partial t^2} \quad \Rightarrow \quad \frac{\partial^2 (af_1+bf_2)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 (af_1+bf_2)}{\partial t^2} \tag{1.1.16}$$

What this tells us is that the wave equation assures the basic wave features, but more information (commonly referred to as *boundary conditions*) is required to get the *specific* wave for the physical situation at hand. It should also be noted that not just any waves work in this way. The waves must both satisfy the same wave equation, which means, for example, that they must have the same speed (though

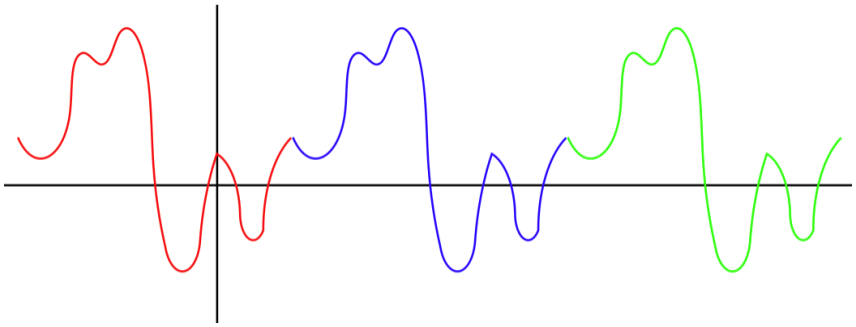
interestingly – and as we will see, importantly – they can be moving in opposite directions, since the velocity appears as a squared value in the wave equation). Also, a one-dimensional wave function cannot be added to a two-dimensional wave function.

1.2: Wave Properties

Periodic Waves

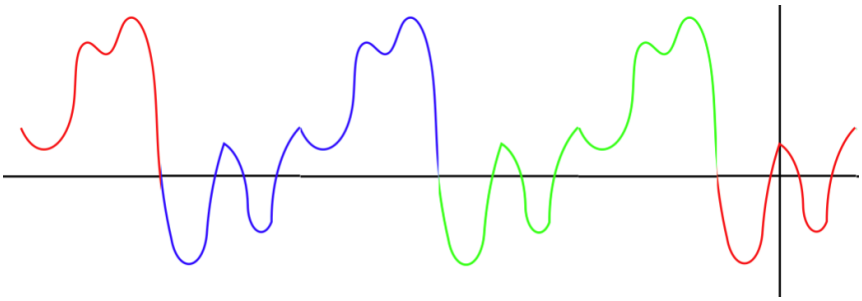
There are qualities that are not required of general waves which are nonetheless common features of waves encountered in nature. The most common special characteristic of a wave is when it continually repeats a specific waveform as it propagates. Such a wave is said to be *periodic*. There are a couple ways to determine if a wave is periodic. The first is to take a snapshot of the wave, and see if its waveform is repeated in space:

Figure 1.2.1a – Snapshot of a Periodic Wave



It should be noted that the starting point of each waveform in the diagram above was chosen arbitrarily. That is, if we look at the same snapshot of the wave as above, we could just as easily demonstrate its periodic nature with different segments:

Figure 1.2.1b – Snapshot of a Periodic Wave



The second way to determine if a wave is periodic is mathematical. The function repeats itself upon translation by a certain distance in the $\pm x$ direction. That is:

$$f\left(x \pm vt\right) = f\left(x \pm vt \pm n\lambda\right), \qquad n = 0,\ 1,\ 2, \ldots \tag{1.2.1}$$

The quantity λ is the length of the repeating waveform, and is called the *wavelength* of the wave. A glance at the two diagrams above should make it clear that the wavelength is a universal feature of that particular wave, and does not depend upon where we choose the starting point to be.

The snapshot of the wave tells us something about its spatial features, but the wave is moving, so if we want to know something about its time-dependence, we need to select a specific point in space, and observe the displacement of the medium as the wave goes by. The wave moves at a constant speed, and the length of each repeating waveform is the same, so the time span required for a single waveform to go by is a constant for the entire wave, called the *period* of the wave. An alternative way of measuring the temporal feature of the wave is the rate at which medium displacements repeat, called *frequency*. Frequency is measured in units of cycles per second, a unit known as *hertz* (*Hz*). Since 1 period is the time required for one cycle, there is a simple relationship between these quantities:

$$f = \frac{1}{T} \tag{1.2.2}$$

We can make another association of periodic wave properties. If we pick a specific point on a waveform (called a *point of fixed phase* for the wave), and follow its motion, it should be clear that it travels a full wavelength in the time of one period. We

therefore can relate the wave speed, wavelength, and period (or frequency):

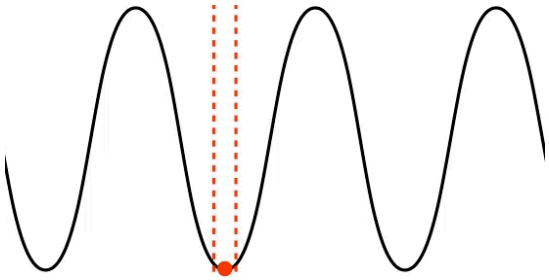
$$v = \frac{\lambda}{T} = \lambda f \tag{1.2.3}$$

Wave Polarization

While the disturbance is not always a displacement of a medium, it always has a directional element to it. A wave that actually displaces a medium has an obvious direction: that of the displacement. Other waves have directional gradients that signify a direction. The direction in which the pressure is changing fastest (the pressure gradient direction) defines a direction for sound waves, and the direction of the electric field vectors defines a direction for light. This directional aspect of waves is also given a name: *polarization*. Generally the direction of medium displacement or gradient is compared to the direction of the wave's motion. There are two special cases that we will encounter for polarization of a wave:

transverse polarization: the medium's displacement or gradient is perpendicular to the wave's direction of motion

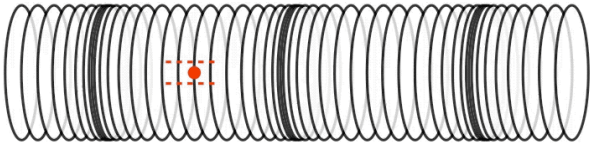
Figure 1.2.2 – Transverse Wave



Note that the displacement of a single point in the medium (depicted by the red dot) is moving only vertically, while the wave moves horizontally. That these two motions are perpendicular to each other is the defining characteristic of a transversely polarized wave. Waves on strings and surface water waves are examples of this kind of wave. As noted earlier, not all waves involve the medium displacing (we will see some examples where this is the case later), but whatever fluctuation is occurring has a direction that can be compared with the direction of the wave's motion.

longitudinal polarization: the medium's displacement or gradient is parallel to the wave's direction of motion.

Figure 1.2.3 – Longitudinal Wave



This time the displacement of a single point in the medium is parallel to the direction of the motion of the wave, the defining characteristic of a longitudinally polarized wave. Notice that like any other wave, the medium is not traveling with the wave, it is moving back-and forth. Physically these are waves induced by *compressions* (regions where the medium is more dense) and *rarefactions* (regions where the medium is less dense). These kinds of waves can be created in springs (as depicted above), but the most common physical example of this kind of wave is sound. Any medium (solid, liquid, or gas) will react to compression, and will therefore exhibit this kind of wave.

Alert

Snapshot graphs of waves of both kinds of polarization are sketched graphically with the displacement on the vertical axis and the position on the horizontal axis. When this is done, it "looks like" a transverse wave, but it is important to keep in mind that such a graph is not a picture of the wave. The vertical axis measures the displacement of the medium from the equilibrium point, which in the case of the red dot on the spring coil for the longitudinal wave in Figure 1.2.3 is the center of the horizontal dotted red lines.

Harmonic Waves

In the category of periodic waves, the easiest to work with mathematically are *harmonic waves*. The word "harmonic" is basically synonymous with "sinusoidal." For a one-dimensional wave, one might therefore assume that a harmonic wave function looks like:

$$f(x,t) = A\cos(x \pm vt) \tag{1.2.4}$$

While this comes close, it has a problem with units. The *total phase* of the wave function (the part in parentheses that is the argument of the cosine) cannot have any physical units, and this function has a phase with units of length. We can therefore repair this problem by dividing the phase by a constant of the wave that has units of length. The obvious such constant is the wavelength. So now our candidate wave function is:

$$f(x,t) = A\cos\left[\frac{1}{\lambda}(x \pm vt)\right] \tag{1.2.5}$$

This gets close, but if we are using radians as the measurement of phase, there is one more change we must add. If we consider a snapshot of this wave at $t = 0$, we would find that the sinusoidal waveform should repeat itself every time the value of x is displaced by λ. If we are using radians as our angular measure, then this requires multiplying the phase by 2π. Then every change of x by λ will result in a change in the phase by 2π, and the function repeats itself properly. So we now have:

$$f(x,t) = A\cos\left[\frac{2\pi}{\lambda}(x \pm vt)\right] \tag{1.2.6}$$

There is one final addition to the phase that we need to make. Suppose we take a snapshot of the wave at $t = 0$ and look at the origin, $x = 0$. This function tells us that the value of the wave's displacement must be its maximum: A. This is not a very general wave! To account for the possibility that the wave might have a different initial condition at the origin, we need to include a *phase constant*, ϕ. Distributing the factor of $\frac{2\pi}{\lambda}$, and using Equation 1.2.3, we get the final form of the wave function of a 1-dimensional harmonic wave:

$$f(x,t) = A\cos\left(\frac{2\pi}{\lambda}x \pm \frac{2\pi}{T}t + \phi\right) \tag{1.2.7}$$

It is common to write this wave function in more compact ways. The first involves the definition of the *wave number* k, and *angular frequency* ω:

$$k \equiv \frac{2\pi}{\lambda}, \quad \omega \equiv 2\pi f = \frac{2\pi}{T} \quad \Rightarrow \quad f(x,t) = A\cos(kx \pm \omega t + \phi) \tag{1.2.8}$$

Another definition that saves even more space is lumping the total phase of the wave into a single function variable: $\Phi(x,t)$. It is clearly linear in the variables x and t. That is:

$$f(x,t) = A\cos(\Phi), \qquad \Phi(x,t) = \frac{2\pi}{\lambda}x \pm \frac{2\pi}{T}t + \phi = kx \pm \omega t + \phi \tag{1.2.9}$$

Finally, it should be noted that although the cosine function was arbitrarily chosen here, we could have just as easily chose a sine function. The only difference between representing the wave with these two functions is the phase constant. That is, we can change from one function to the other if we change the phase constant by $\frac{\pi}{2}$:

$$\cos(\Phi) = \sin\left(\Phi + \frac{\pi}{2}\right) \quad \Rightarrow \quad \phi \to \phi + \frac{\pi}{2} \tag{1.2.10}$$

Separation of Variables

The important thing to take away from the harmonic wave function in Equation 1.2.7 is that the wave has four *constants of the motion* that completely define it. Besides the wavelength, period, and phase constant, there is the *amplitude*, A. All of these remain fixed in time, completely defining the wave that evolves thanks to its x and t dependence. These constants can be extracted from what was referred to in the previous section as boundary conditions.

It turns out that harmonic wave functions have another feature that makes them special. To see this, let's employ a powerful method that is used to solve equations like the wave equation called *separation of variables*. We will not go into great detail here (you will see this used later in this course, and will see it over and over in future math and physics courses), but you will get a feel for how it works. We will stick to the one-dimensional wave equation to keep it as simple as possible.

It goes like this: Let's look for solutions to the wave equation that satisfy a very specific criterion – let's assume the wave function can be separated into a product of two functions, one of them a function only of position, and the other a function of only time:

$$f(x,t) = \mathcal{X}(x) \cdot \mathcal{T}(t) \tag{1.2.11}$$

It should be immediately clear that only a special group of wave functions will satisfy this condition. For example, $f(x-vt) = (x-vt)^2$ is a one-dimensional wave function that cannot be written as a product of two functions with one of them only a function of x and the other only a function of t. Let's plug our special wave function into the wave equation:

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{1}{v^2}\frac{\partial^2}{\partial t^2} f(x,t) \quad \Rightarrow \quad \frac{\partial^2}{\partial x^2}[\mathcal{X}(x) \cdot \mathcal{T}(t)] = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}[\mathcal{X}(x) \cdot \mathcal{T}(t)] \tag{1.2.12}$$

Because of the nature of partial derivatives, the left side of this equation only takes two derivatives of the function $\mathcal{X}(x)$, and the right side only two derivatives of the function $\mathcal{T}(t)$. These derivatives are of single-variable functions, so we don't even need to refer to them as partial derivatives when they act on the partial functions. Representing ordinary derivatives with primes, we have:

$$\mathcal{X}''(x) \cdot \mathcal{T}(t) = \frac{1}{v^2}\mathcal{X}(x) \cdot \mathcal{T}''(t) \tag{1.2.13}$$

Dividing both sides of this equation by $\mathcal{X}(x) \cdot \mathcal{T}(t)$ gives:

$$\frac{\mathcal{X}''(x)}{\mathcal{X}(x)} = \frac{1}{v^2}\frac{\mathcal{T}''(t)}{\mathcal{T}(t)} \tag{1.2.14}$$

And now for the magic... Notice that the left side of this equation is only a function of x, while the right side is only a function of t (remember, v is a constant for the wave, and doesn't depend on either x or t). The variables x and t are completely independent of each other – we can look at different positions on the wave at the same time, or at a single position on the wave at different times. For two functions of independent variables to be equal to each other, it means that they must both equal the same constant. For example, if the ratio on the left side of the equation depended on x, then the right side would have to be a function of x, and it is not. Expressing this fact mathematically (and choosing a form of the constant that will make sense later), we have:

$$\frac{\mathcal{X}''(x)}{\mathcal{X}(x)} = \frac{1}{v^2}\frac{\mathcal{T}''(t)}{\mathcal{T}(t)} = -k^2 \tag{1.2.15}$$

We can now write two separate equations, one exclusively in terms of x, and the other exclusively in terms of t (thus the name of this method!):

$$\mathcal{X}'' + k^2\mathcal{X} = 0 \qquad \mathcal{T}'' + k^2v^2\mathcal{T} = 0 \tag{1.2.16}$$

These two differential equations are the same – they both involve two derivatives of a function equaling a constant times the function again. What is more, we have seen this differential equation before! We know that either a sine or a cosine function will satisfy these equations, so a linear combination will as well. We can therefore write solutions to these two equations as:

$$\mathcal{X}(x) = a\cos kx + b\sin kx \qquad \mathcal{T}(t) = c\cos\omega t + d\sin\omega t \tag{1.2.17}$$

[*Here we have defined* $\omega \equiv kv$.] So now we can reconstruct the full wave function:

$$f(x,t) = \mathcal{X}(x) \cdot \mathcal{T}(t) = (a\cos kx + b\sin kx)(c\cos\omega t + d\sin\omega t) \tag{1.2.18}$$

What this result says is that any wave of the given one-dimensional wave equation that can be separated into a product of two single-variable functions can be written in this form, and the specifics of that wave are given by the constants a, b, c, d, k, and ω. But there is one important restriction to keep in mind: The ratio $\frac{\omega}{k}$ must equal the speed of the wave in question.

It is interesting to note that the basic cosine wave function given above in Equation 1.2.8 is *not* a separable solution to the wave equation. That is, one cannot reach Equation 1.2.8 (for given values of k and ω) by an appropriate choice of the constants a, b, c, and d. It can only be reached using a linear combination of two different separable solutions. We will cover this case in detail in a later chapter.

1.3: Energy Transmission

1-Dimensional Waves

While we will be interested in energy transmission in all kinds of waves, we will start with 1-dimensional harmonic waves as our model, as we are already familiar with the harmonic motion exhibited by the medium as such a wave passes. In particular, we have an idea of how to deal with the *energy* of a single oscillating particle, so for now we will also restrict ourselves to mechanical waves, where the particle comprising the medium are actually oscillating (i.e. think about a transverse wave on a string).

The energy of a single oscillating particle comes in two forms: kinetic and elastic potential (we'll maintain the convention that the particle is displacing in the y direction as the wave moves in the $\pm x$ direction):

$$E_{tot} = KE + PE_{elastic} = \frac{1}{2}mv^2 + \frac{1}{2}ky^2 \tag{1.3.1}$$

Of course both the speed of the particle and its displacement are changing with time, so it's more useful to express the energy of this particle in terms of one of the constants of the motion. When the particle reaches its maximum displacement, it stops moving, so its kinetic energy goes to zero and all of the energy is potential. But we have given this maximum displacement a name – amplitude. So the total energy of the oscillating particle is:

$$E_{tot} = \frac{1}{2}kA^2 \tag{1.3.2}$$

One might complain that there are no springs present for this kind of wave, so what are we supposed to plug into k? Well, there *is* a restoring force on every particle in the string as the wave passes, and this *behaves* like the restoring force of a spring, but we can write this expression more appropriately if we replace the spring constant with an equivalent expression in terms of the mass of the particle and the frequency of oscillation. Recall that for simple harmonic motion we have:

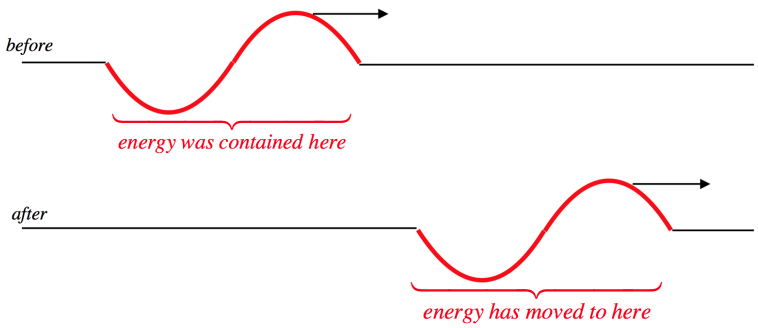
$$2\pi f = \omega = \sqrt{\frac{k}{m}} \quad \Rightarrow \quad k = m(2\pi f)^2 \tag{1.3.3}$$

Now the energy of the particle is in terms of the medium (the mass of the particle) and the wave (the amplitude and frequency):

$$E_{tot} = \frac{1}{2}m(2\pi f)^2 A^2 \tag{1.3.4}$$

We stated at the very beginning that waves carry energy from one point to another. Now that we see that a single particle in the medium carries energy, it should be clear that this is true. Consider a wave pulse that is harmonic for just one wavelength:

Figure 1.3.1 – Wave Pulse Carries Energy



Clearly the region where the particles are oscillating changes in this case, which means that the region that contains the energy is changed. The pulse transports energy across the expanse by having particles in the medium transfer energy to their nearest-neighbors, without the particles themselves having to make the trip.

Suppose we wish to know how much energy is in the whole wave, rather than what is just in a single particle. In this case, we treat the wave as continuous, with an infinite number of infinitesimal particles oscillating. The mass of these particles is very small, and can be written in terms of the mass density of the medium (again, think of this as a wave on a string), multplied by a small segment along the x direction:

$$dm = \mu dx \tag{1.3.5}$$

We can now use this mass to express the infinitesimal amount of energy possessed by that particle, by plugging this into Equation 1.3.4:

$$dE = \frac{1}{2} dm (2\pi f)^2 A^2 = \frac{1}{2} \mu dx (2\pi f)^2 A^2 \tag{1.3.6}$$

This is the energy in a single particle of the medium within the wave, so to get the full energy carried by the wave, we need only add up all these parts by performing an integral. The range of the single wave goes for one wavelength, so choosing the origin to be at one end of the wave, we have:

$$E_{in\ wave} = \int\limits_{in\ wave} dE = \int\limits_0^\lambda \frac{1}{2} \mu dx (2\pi f)^2 A^2 = \frac{1}{2} \mu (2\pi f)^2 A^2 \int\limits_0^\lambda dx = \frac{1}{2} \mu (2\pi f)^2 A^2 \lambda \tag{1.3.7}$$

Of course, if we have a full harmonic wave, as described by the wave function given in Equation 1.2.7, we have an infinite number of these single wave pulses, and the *amount* of energy in the entire wave is infinite and uninteresting. What is finite, even in the case of a full harmonic wave, is the *rate* as which energy is being transferred. To compute this, we simply need to divide how much energy a single wave pulse is carrying by the time it takes to completely cross some fixed point. Well, we know that this time interval is one period, so we have for the power of the wave:

$$P_{wave} = \frac{E_{one\ wave\ pulse}}{T} = \frac{1}{2} \mu (2\pi f)^2 A^2 \frac{\lambda}{T} = \frac{1}{2} \mu (2\pi f)^2 A^2 v, \tag{1.3.8}$$

where v is the speed of the wave.

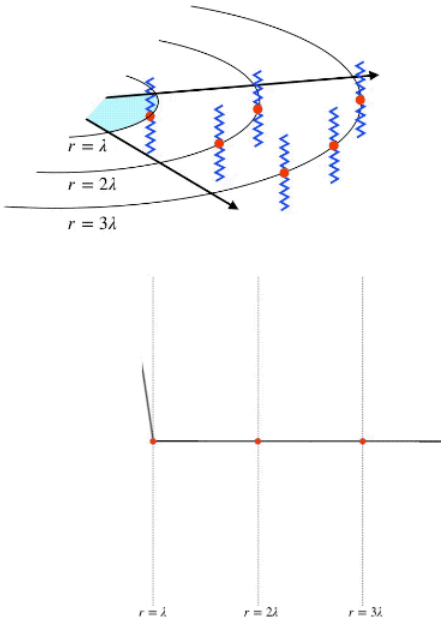
This calculation is specific to harmonic waves on strings, and we will not go into how this result changes for other types of harmonic waves (which pass through different sorts of media, may not be mechanical in nature, etc.). However, we will note that for a one-dimensional wave, the power is proportional to the square of the amplitude. As we will see, we will need to modify this result slightly for waves in two and three dimensions.

Multi-Dimensional Waves

In Section 1.1 we found that in order to satisfy the wave equation, waves that propagate out from a central source, into two or three dimensions cannot repeat their waveform. Here we will see why that is so, and get some physical idea of specifically how the waveform changes (we already saw that it does mathematically). Again, we will remain within the confines of our harmonic wave model for simplicity. First we need to clarify an important assumption: In our discussion we will assume 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.

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.

Figure 1.3.2 – Circular Wave Energy Conservation



As the wave moves out, there is no energy lost, so the 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. Specifically, the energy per oscillator is proportional to the *square* of the amplitude (Equation 1.3.2), which means that doubling the radius of the circle reduces the amplitude by a factor of $\sqrt{2}$, tripling the radius reduces the amplitude by a factor of $\sqrt{3}$, and so on. 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. 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:

$$I_{2d}(r) = \frac{P}{2\pi r} \tag{1.3.9}$$

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:

$$A \propto \frac{1}{\sqrt{r}} \quad \Rightarrow \quad I \propto A^2 \tag{1.3.10}$$

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:

$$I_{3d}(r) = \frac{P}{4\pi r^2} \tag{1.3.11}$$

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:

$$A \propto \frac{1}{r} \quad \Rightarrow \quad I \propto A^2 \tag{1.3.12}$$

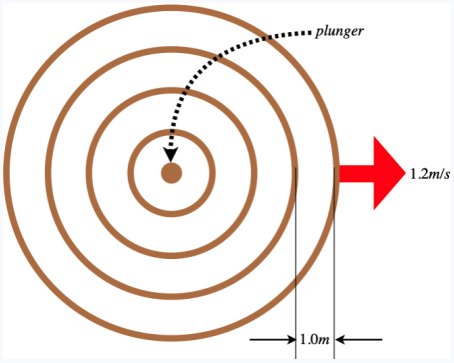
The relation between intensity and amplitude is therefore universal among waves, and one that we will keep in mind in the sections to come.

Note that this intensity drops faster than that of the two-dimensional wave, satisfying what's known as an *inverse-square law*: The intensity gets weaker in inverse proportion to the square of the distance from the source. Since the power of the wave is the same everywhere, we have the following relationship of intensities at two distances r_1 and r_2 from the source for waves that propagate outward from a point source:

$$I_1 r_1^2 = I_2 r_2^2 \tag{1.3.13}$$

Example 1.3.1

At time $t = 0$, a plunger begins oscillating up-and-down at a steady rate for 6 full oscillations in a body of otherwise calm water. During this time, it puts $420J$ of energy into the surface waves it creates. The wavelength of the wave is measured to be $1.0m$, and the wave speed is measured to be $1.2m/s$.



a. *Find the power supplied by the plunger.*
b. *Find the intensity of the leading wavefront at time $t = 4.0s$.*
c. *The amplitude of the leading wavefront at $t = 4.0s$ is measured to be $5.4cm$. Find its amplitude at $t = 8.0s$.*

Solution

a. The power is the rate at which the energy is being transferred into the waves. We know how much energy is put into 6 oscillations, so if we divide that energy by the time span of 6 oscillations, we have the value of the power. The time span of 6 oscillations is 6 periods, and a single period we can calculate from the wavelength and wave speed:

$$T = \frac{\lambda}{v} = \frac{5}{6}s \quad \Rightarrow \quad \Delta t = 6T = 5.0s \quad \Rightarrow \quad P = \frac{E}{\Delta t} = \frac{420J}{5s} = 84W$$

b. To get the intensity, we need to know the circumference of the leading wavefront. We know the speed of the wave and how long it has been traveling, so:

$$r = v\Delta t = \left(1.2\frac{m}{s}\right)(4.0s) = 4.8m \quad \Rightarrow \quad I = \frac{P}{2\pi r} = \frac{84W}{2\pi 4.8m} = 2.8\frac{W}{m}$$

c. For the two-dimensional wave, the amplitude gets smaller as the radius grows, by a factor of $\frac{1}{\sqrt{r}}$. The wave's speed is unchanging, so after 8 seconds the wave has traveled twice as far from the source than after 4 seconds. Doubling the distance traveled therefore reduces the amplitude by a factor of $\sqrt{2}$, giving:

$$A\,(t = 8.0s) = \frac{A\,(t = 4.0s)}{\sqrt{2}} = \frac{5.4cm}{\sqrt{2}} = 3.8cm$$

1.4: Superposition and Interference

Combining Similar Waves

When two or more waves of the same type in the same medium coexist in the same region of space, they combine to create a new wave. The way they combine is a simple process known as *superposition*. This consists of simply adding the displacements (or whatever the wave function represents) of the two or more waves at the same place and time. For a 1-dimensional wave, this means:

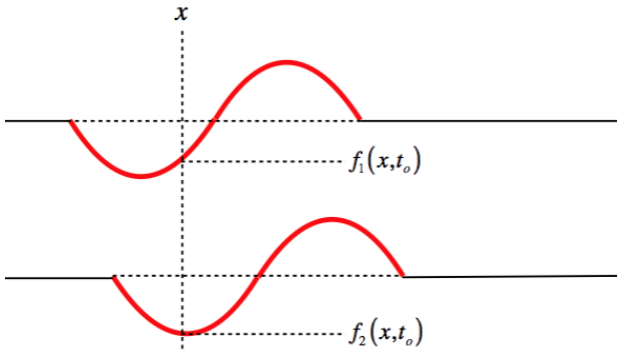
$$f_{tot}(x,t) = f_1(x,t) + f_2(x,t) \tag{1.4.1}$$

Alert

It's important to emphasize that two waves can only superpose if they are the same type. Many different kinds of waves can travel through the same medium (light, sound, and displacement waves can all travel through water in a lake, for instance), but these cannot superpose with each other.

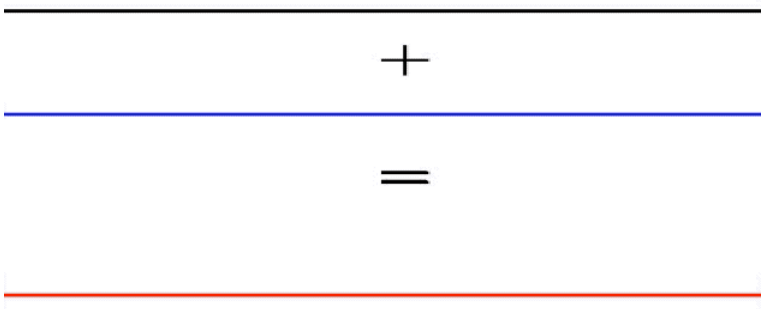
We showed in Equation 1.1.16 that if the two individual wave functions satisfy the wave equation, then so does the total wave function. It bears repeating with a diagram that this superposition sum involves adding displacements *at the same place and time*. So if we took a snapshot of two waves, we would determine the total wave by lining them up:

Figure 1.4.1 – Superposition



The composite wave is then the combination of all of the points added thus. Of course, these are traveling waves, so over time the superposition produces a composite wave that can vary with time in interesting ways. Here is a simple example of two pulses "colliding" (the "sum" of the top two waves yields the bottom wave).

Figure 1.4.2 – "Collision" of Pulses



Notice that even though the resultant wave looks very different from its "parents," the medium somehow "remembers" the original waves, and when they no longer coincide, they continue along as exactly the waves they were before the superposition. That is, the waves do not affect each other, as particles would if they collided – waves don't bounce off each other, for example. They simply

create a new wave while they occupy the same space in the medium, and when their individual motions carry them to different parts of the medium, they return to being the waves they were before.

Interference – All-or-Nothing

There are some special cases involving superposition that are particularly interesting to examine, and these involve a phenomenon known as *interference*. There are many degrees of interference possible, all of which fall between the following two extremes:

- *constructive interference*: The waves are perfectly aligned and timed so that their crests and troughs coincide, such that the total wave has the maximum possible amplitude (equal to the sum of the amplitudes of the two constituent waves).
- *destructive interference*: The waves are perfectly aligned and timed so that the crests of one wave align with the troughs of the other such that leading to a wave that has the minimum possible amplitude (equal to the difference of the amplitudes of the two constituent waves).

The phrase *total destructive interference* refers to the case of destructive interference when the resultant wave has zero amplitude, i.e. the two waves totally cancel each other. In the cases we will discuss, we will only talk about this extreme case of destructive interference, so we will typically leave out the word "total," even though we are still talking about total cancelation.

Interference – Intensity of Combined Wave

We will examine a great many examples of interference in physical phenomena in the sections to come. We therefore need to take some time to develop the mathematics behind this effect. We will do this within the same framework that we have been using – that of harmonic waves. When we look at the physical attributes of interference, what we will be examining is what happens to the intensity of the combined wave. For example, interference in sound will be exhibited in volume, and in light it will be brightness. Both of these are measures of intensity. We need a reference point for intensity, and the one we will use is that of maximal constructive interference. So what we seek is an equation that relates the intensity of two superposed, out-of-phase, but otherwise identical waves to the intensity we would see if they were in phase. That is, we want something that looks like this:

$$I\left(\Delta\Phi\right) = I_o g\left(\Delta\Phi\right) \tag{1.4.2}$$

The quantity I is the intensity of the wave as a function of the phase difference of the two (identical) parent waves. If the two waves happen to be in phase, then the combined wave's intensity is I_o when the two waves are in phase. Note that this is *four times the intensity of each individual wave*, since the constructive interference adds the amplitudes (which are equal – the waves are identical) and the intensity is proportional to the square of the amplitude.

The function needs to have the following properties:

- It has to always be non-negative, since intensity is never a negative number.
- It has to vanish when the phase difference equals π (modulo 2π), since this means the waves totally destructively interfere.
- It has to equal 1 when when the phase difference is 0 (modulo 2π), since this means the waves constructively interfere.

To find this function, we start with two wave functions that are identical except for their phases and superpose them:

$$f_{tot} = f_1 + f_2 = A\cos(\Phi_1) + A\cos(\Phi_2) = A\left[\cos(\Phi_1) + \cos(\Phi_2)\right] \tag{1.4.3}$$

We want this function to only depend upon the difference in the two phases, so we will write each total phase in terms of deviation from their average phase (which we will call simply Φ), and the difference in phase between the two waves, $\Delta\Phi$:

$$\Phi = \frac{\Phi_1 + \Phi_2}{2} \quad \Rightarrow \quad \Phi_1 = \Phi - \frac{\Delta\Phi}{2}, \quad \Phi_2 = \Phi + \frac{\Delta\Phi}{2} \tag{1.4.4}$$

Plug these into Equation 1.4.3 gives:

$$f_{tot} = A\left[\cos\left(\Phi - \frac{\Delta\Phi}{2}\right) + \cos\left(\Phi + \frac{\Delta\Phi}{2}\right)\right] \tag{1.4.5}$$

Now we can apply a trigonometric identity:

$$\cos(A-B) + \cos(A+B) = 2\cos A\cos B \quad \Rightarrow \quad f_{tot} = 2A\cos(\Phi)\cos\left(\frac{\Delta\Phi}{2}\right) \tag{1.4.6}$$

The phase difference between the two waves can be written in terms of the difference in position, time, and the phase constant, using Equation 1.2.9:

$$\Delta\Phi = \frac{2\pi}{\lambda}(x_1 - x_2) \pm \frac{2\pi}{T}(t_1 - t_2) + \phi_1 - \phi_2 = \frac{2\pi}{\lambda}\Delta x \pm \frac{2\pi}{T}\Delta t + \Delta\phi \tag{1.4.7}$$

As the waves propagate along, the values of x and t will change, but as the two waves are identical (traveling in the same direction with the same speed), the differences in x and t don't change for a given phase. Therefore the factor in Equation 1.4.6 that includes the phase difference is a constant. Putting that constant together with the $2A$ gives us the amplitude of the new conglomerate wave (with the time-varying phase being the average of the phases of the two waves):

$$f_{tot} = A_{new}\cos(\Phi), \qquad A_{new} \equiv 2A\cos\left(\frac{\Delta\Phi}{2}\right) \tag{1.4.8}$$

The intensity is proportional to the square of the amplitude, so the intensity of this combined wave is:

$$I \propto A_{new}^2 = 4A^2\cos^2\left(\frac{\Delta\Phi}{2}\right) \tag{1.4.9}$$

The intensity of each individual wave is proportional to A^2. If the waves were in phase, the total amplitude would double, which means that the total in-phase intensity I_o is proportional to $4A^2$. The intensity of the (out of phase) combined wave is therefore:

$$I = I_o\cos^2\left(\frac{\Delta\Phi}{2}\right) \tag{1.4.10}$$

Notice that this relationship between total intensity and phase difference exactly matches the three criteria we outlined above.

How to Create Interference

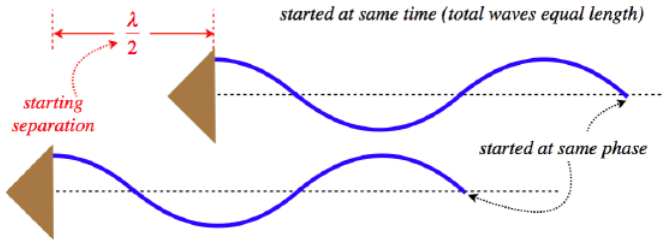
Whenever two waves interfere, whether it is constructively, destructively, or anything in-between, it's clear that the critical factor is the phase difference between the waves, $\Delta\Phi$. From Equation 1.4.5 we can see several ways in which a difference in phase can occur: The two waves can travel a different distance (Δx), they can have been traveling for a different amount of time (Δt), they could have started out of phase ($\Delta\phi$), or it could be some combination of these three differences.

To understand how these three differences can be physically manifested, it's easiest to let two of them be zero, and let only one difference occur at a time. We can do this in two ways. The first is to simply construct physical situations that assures this, and the second consists of nothing more than a change of perspective. For the sake of studying this effect, we will only consider destructive interference, but we can do the same for constructive or anything in between as well. To keep things simple, we'll interfere (approximately) 1-dimensional waves traveling in the same direction, which we can model with identical harmonic sound waves coming from two speakers pointed in the same direction. There is nothing about the result that is specific to sound, however; this is a phenomenon common to all waves.

Case 1: Different Travel Distance ($\Delta x \neq 0$, $\Delta t = 0$, $\Delta\phi = 0$)

We start with a case where the two sound waves emanate from their respective speakers such that the leading wave front for each sound wave is at the same phase (in the diagram below, if we use a cosine function to describe these waves, then both waves have a phase of $\frac{\pi}{2}$ at their leading edge). We also start the waves from their speakers at the same moment (in the diagram below, both waves have been propagating for one full oscillation plus another quarter of an oscillation, so they started at the same time). But the waves are offset in the positions where they begin by one-half wavelength, which results in the two waves occupying the same medium π radians out of phase:

Figure 1.4.3 – Destructive Interference Due to Travel Distance Only

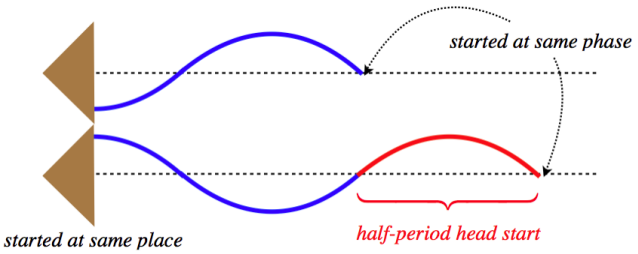


$$\Delta\Phi = 2\pi \underbrace{\Delta\left(\frac{x}{\lambda}\right)}_{\frac{1}{2}} \pm 2\pi \underbrace{\Delta\left(\frac{t}{T}\right)}_{0} + \underbrace{\Delta\phi}_{0} = \pi \quad (\textit{destructive interference})$$

Case 2: Different Time Intervals $(\Delta x = 0, \Delta t \neq 0, \Delta\phi = 0)$

Next we will place the speakers side-by-side, so that the travel distances of the waves at any position in the medium is the same. As before, the leading wave front of the waves will be in phase, but this time we will turn on one of the speakers at a time of one-half period before the other speaker. This lag between the two once again throws the two waves out of phase, resulting in destructive interference:

Figure 1.4.4 – Destructive Interference Due to Starting Time Only

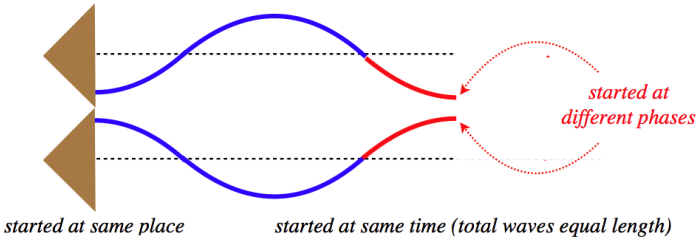


$$\Delta\Phi = 2\pi \underbrace{\Delta\left(\frac{x}{\lambda}\right)}_{0} \pm 2\pi \underbrace{\Delta\left(\frac{t}{T}\right)}_{\frac{1}{2}} + \underbrace{\Delta\phi}_{0} = \pi \quad (\textit{destructive interference})$$

Case 3: Different Starting Phases $(\Delta x = 0, \Delta t = 0, \Delta\phi \neq 0)$

Now finally, we will position speakers side-by-side, and turn them on at the same moment, but will arrange things so that the sounds they emanate leave the speakers π radians out of phase, to get the same destructive interference:

Figure 1.4.5 – Destructive Interference Due to Phase Constant Only



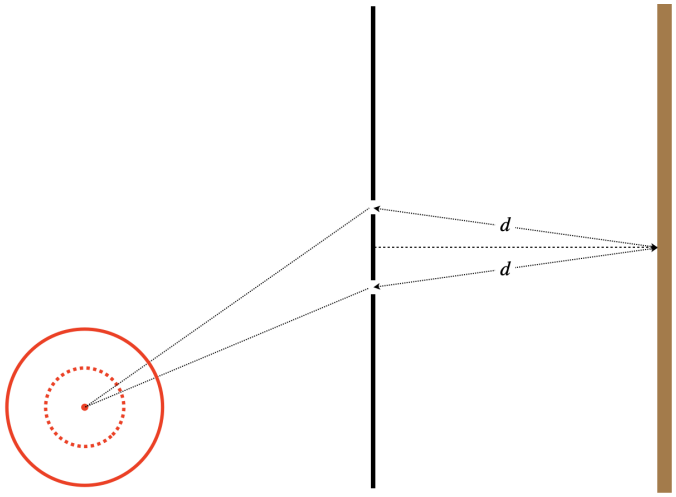
$$\Delta\Phi = 2\pi \underbrace{\Delta\left(\frac{x}{\lambda}\right)}_{0} \pm 2\pi \underbrace{\Delta\left(\frac{t}{T}\right)}_{0} + \underbrace{\Delta\phi}_{\pi} = \pi \quad (\textit{destructive interference})$$

One More Case: Different Perspectives

While these all seem distinctly different, very often the same interference effect can be described in any of the three ways, simply by changing our perspective. To see this, let's consider a two-dimensional harmonic wave coming from a single point source. This could be a wave caused by a pebble dropped into a pond, for example. This wave radiates circularly-outward from the source. We of course cannot have interference with only one wave, so we will provide a means to split it into two separate waves: The wave will strike a barrier with two small holes in it, through which the wave can pass. As we will see in a future section, these two holes themselves now act like point sources of two separate waves, with the energy of these waves coming from the original wave. It's these two waves that we will allow to interfere.

Consider the diagram below. We will be looking at the intensity of the wave when it strikes a second barrier at a position that is equal distances (labeled as 'd' in the diagram) from the two holes. The original source of the wave is *not* the same distance from both holes, however.

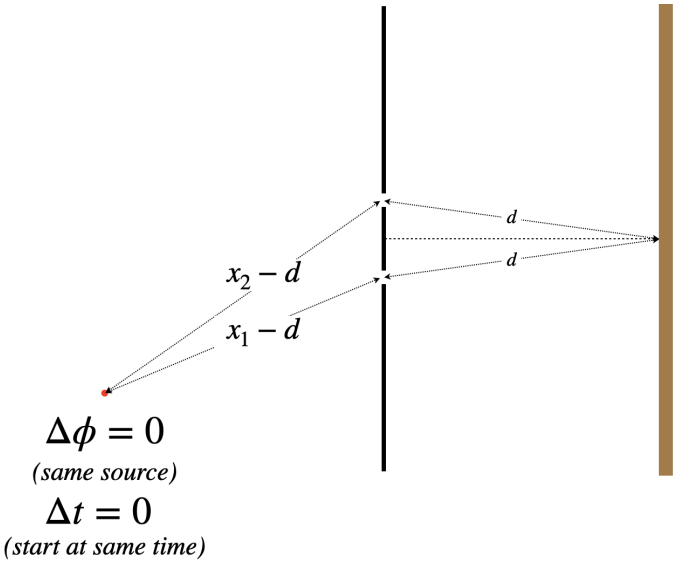
Figure 1.4.6 – Single Source Interference Setup



Suppose we find that the two split waves interfere destructively at the final destination. How can we explain this result? It turns out that we can do it in any of the three ways described above, depending on what perspective we decide to take.

First, we can note that the two waves start at the same time and with the same phase, from the original point source. But these two waves (which initially are part of the same starting wave) do *not* travel equal distances: $x_1 \neq x_2$. They travel the same distance from the holes to the screen, but the distances to the holes from the starting point are different:

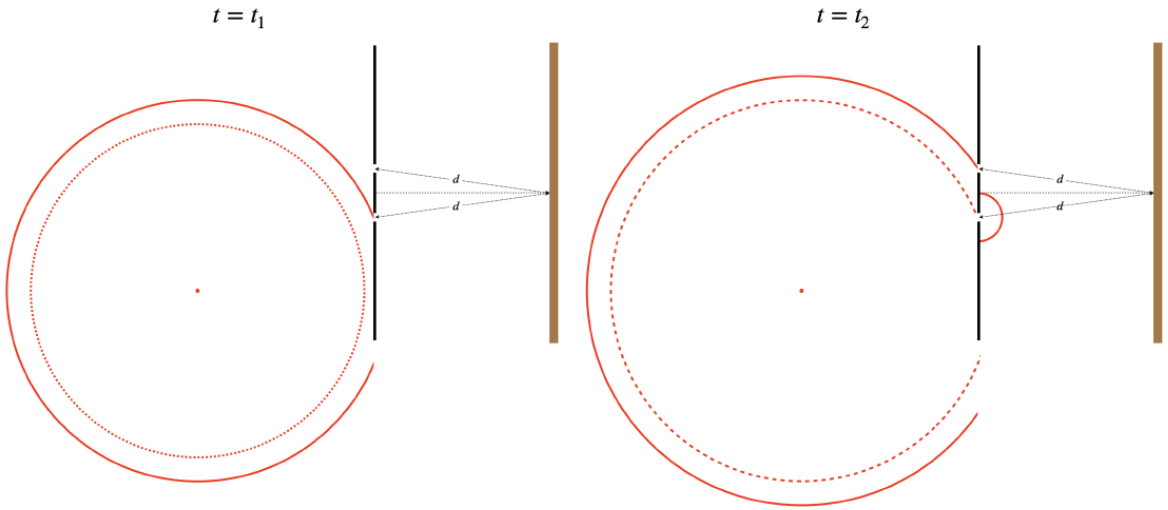
Figure 1.4.7 – Single Source Interference – Distance Traveled Explanation



We would see destructive interference if (for example), the extra distance traveled by the wave passing through one hole happens to be a half wavelength farther than the distance traveled by the wave passing through the other hole: $x_2 - x_1 = \frac{1}{2}\lambda$.

Now let's change our perspective. Suppose we are watching this wave from the right of the two holes, and know nothing at all about the single point source. As far as we are concerned, the two waves are starting from positions that are the *same distance* from the position where we are seeing the destructive interference. We would not conclude that the difference in distance traveled is the cause of this interference. But suppose we had been watching since before the first wave emerged from a hole. In this case, we would see a wave with a certain starting phase come out of one hole first, and a short time later, a wave with the same phase come from the other hole. It comes out with the same phase because it comes from the same wavefront from the original point source.

Figure 1.4.8 – Single Source Interference – Time Elapsed Explanation

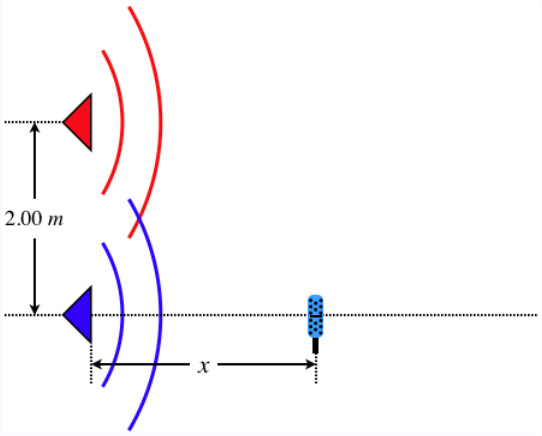


We would see destructive interference if (for example), the time elapsed between when we see the two waves emerge differs by one-half period: $t_2 - t_1 = \frac{1}{2}T$.

And finally, there is one other perspective from which we can view this. Once again, we will view the waves from the side of the holes where we can't see the point source, so that we again measure equal travel distances. But this time, let's assume we don't start viewing until after the wave has been passing through both holes for awhile. We look at our watch and note that at time $t = 0$ (when we start watching), there are waves coming from both holes that are out of phase with each other by π radians. Both waves travel the same distance and start at the same time, but start out out of phase, and therefore destructively interfere.

Example 1.4.2

Two speakers, both pointing in the $+x$ direction, are placed on the y-axis, separated from each other by a distance of $2.00\ m$. They emit the same tone, which has a frequency of $784\ Hz$, in phase with each other. A microphone is placed directly in front of and very close to one of the speakers, and is gradually moved from along the x-axis farther and farther from the speaker. Assume that the fact that the microphone is a little farther from one speaker than the other does not result in a noticeable intensity difference between the two speakers, so that the sound waves coming from the speakers have the same amplitude when they reach the microphone. The speed of sound waves in air is $344\frac{m}{s}$.



a. Find the distances from the closer speaker where the microphone detects no sound.
b. Find the distances from the closer speaker where the sound gets loudest (i.e. constructive interference).
c. Suppose the tone coming from the speakers has an adjustable frequency, and that it is gradually lowered. Find the frequency below which the microphone has no position on the x-axis where it measures total silence.

Solution

a. The starting phase and time are the same, so the only source of phase difference comes from the difference in distance traveled. From the Δx contribution to the phase difference that causes destructive interference (i.e. when the extra distance is an odd number of half-wavelengths), we therefore have:

$$\frac{2\pi}{\lambda}\Delta x = n\pi \quad \Rightarrow \quad \Delta x = \frac{1}{2}n\lambda\,, \qquad n = 1,\ 3,\ 5,\ \ldots$$

The difference in distances traveled by the two waves us found using the Pythagorean theorem, so putting this in above gives:

$$\Delta x = \sqrt{x^2 + (2.00\ m)^2} - x = \frac{1}{2}n\lambda \tag{1.4.11}$$

Calling the n^{th} value of x "x_n" and doing the algebra gives:

$$x_n = \frac{1}{n\lambda}\left(4.00m^2 - 0.250n^2\lambda^2\right)$$

We are given the frequency of the sound, so we can find its wavelength:

$$\lambda = \frac{v}{f} = \frac{344\frac{m}{s}}{784Hz} = 0.439m$$

Notice that although the value of n is not restricted, when it gets too high, the value of x_n will become negative. Plugging in all of the values of n that give positive values of x yields five possible values:

$$x_1 = 9.00m, \quad x_3 = 2.71m, \quad x_5 = 1.27m, \quad x_7 = 0.531m, \quad x_9 = 0.022m$$

b. Constructive interference occurs when the path difference is an even number of half-wavelengths (i.e. some number of full wavelengths). We can get our answer directly from part (a) simply by taking even values of n instead of odd values. Once again, the number of values of n is limited by the restriction that sign of x_n must be positive.

$$x_2 = 4.34m, \quad x_4 = 1.84m, \quad x_6 = 0.858m, \quad x_8 = 0.259m$$

c. The value of Δx clearly gets smaller as x gets larger (the hypotenuse gets closer and closer to equaling the x value as x gets larger), so the largest possible value of Δx is just the separation of the speakers. If the speakers are separated by less than one-half wavelength, then Δx can never get as big as a half wavelength, and no totally destructive interference is possible. These speakers are separated by $2.00\ m$, so the wavelength of the sound must be shorter than $4.00m$ for there to be any instance of total destructive interference. This wavelength corresponds to a frequency of:

$$f = \frac{v}{\lambda} = \frac{344 \frac{m}{s}}{4.00m} = 86Hz$$

Frequencies lower than this create wavelengths that are so long that Δx is never large enough to cause total destructive interference.

1.5: Standing Waves

Interference Patterns

We found that interference occurs between two identical waves, but we didn't mention what the source of two identical waves might be. We will find that most of the time the two waves are actually the *same* wave, where one part of it has been diverted somehow, so that it behaves like a separate wave. When we witness the interference created in such a situation, it is often in the form of an *interference pattern*. This is a recognizable pattern of intensity that repeats itself in space or in time, or in both. We will see lots of these patterns in the sections to come, but as usual we will start with a simple (but important) one-dimensional example of an interference pattern, called a *standing wave*. [*Actually, standing waves occur in 2 and 3 dimensions as well, though we will confine our discussion to those of the 1-dimensional variety.*]

Alert

The moniker "standing wave" puts yet another strain on our definition of what it means to be a wave. It does satisfy the wave equation (as does any superposition of waves), but although the wave equation yields a wave velocity, this waveform does not propagate at all. It is better to think of standing waves as what they are – interference patterns.

All interference patterns are formed from multiple identical waves, and like so many other interference patterns, this is accomplished through multiple versions of the same wave. In the case of the standing wave, these two versions are the result of wave *reflections* off two endpoints. That is, a single wave bounces back-and forth between two endpoints, and as it crosses itself during the journey, the standing wave interference pattern is formed from the superposition – the two waves that are interfering only differ in their directions of motion.

Wave Reflection & Transmission

Before we delve into the details of standing waves, we first need to look at the phenomenon that makes them possible – wave reflection. The mathematics of wave reflection can become quite involved and we will not delve into it here, but the bottom line is that waves reflect off sudden changes in the medium. We have found that the medium is best characterized by the speed of waves that pass through it, and in fact it is correct to say that a wave reflects when it encounters a region of the medium where the wave speed changes.

At this point one might ask why the wave doesn't simply continue in the direction it was going, but at a different speed. It does! But it also reflects. That is, the wave splits into two parts, called the *reflected wave* and the *transmitted wave*. Of course, energy is conserved during this schism, so the energy in the original wave is greater than the energies in either of these waves. The amount of energy that goes to each wave is determined mathematically by a process known as "matching boundary conditions" at the point of reflection, but as mentioned earlier, we will not make a close examination of this process here (this topic is explored in courses on quantum mechanics, such as Physics 9D).

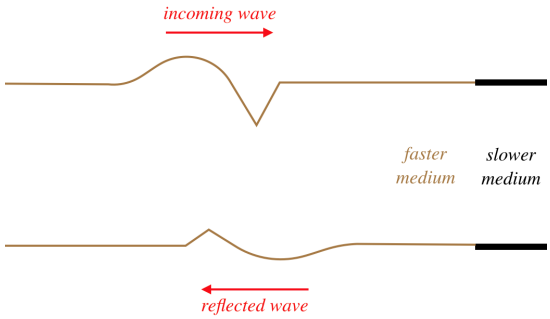
The requirement that the speed of the wave changes at the point of reflection in the medium doesn't distinguish between the wave is coming from a faster medium to a slower one, or from a slower one to a faster one. It turns out that the wave will partially transmit and partially reflect, no matter which direction it is going. But there is an observable phenomenon that distinguishes these two possibilities. Suppose a pulse of a wave on a string consists of just a single bump (like half a sine function) that lies on the top half of the string. If this wave reaches a point in the medium where it speeds up (the string's linear density goes down), then the reflected pulse remains in the top half of the string. But if the pulse encounters a point in the medium where it slows down (the string's linear density goes up), then the reflected pulse flips to the bottom half of the string. The transmitted wave never flips over.

Figure 1.5.1 – Reflection and Transmission (Slow-to-Fast Medium)

Figure 1.5.2 – Reflection and Transmission (Fast-to-Slow Medium)

It should also be noted that the reflected wave in both cases reflects its wave form along $-x \leftrightarrow +x$. This is not obvious in the case of a symmetric pulse, but if the wave is asymmetric, then it becomes apparent.

Figure 1.5.3 – Reflection of an Asymmetric Wave (Fast-to-Slow Medium)



The leading edge of the incoming waveform is the leading edge of the reflected waveform. While there is some loss of amplitude for the reflected wave compared to its incoming counterpart (some of the energy is taken by the transmitted wave), the wavelength of the reflected wave is the same as the incoming wave. This is because the velocity of the reflected wave is the same as the incoming wave. The wavelength of the transmitted wave will not match the wavelength of the incoming wave, however. The time span between the front and rear of the waveform striking the new medium is the same time as it takes for the full waveform to be transmitted, so the periods of the incoming, transmitted, and reflected waves are all the same, but since the velocity is different for the reflected and transmitted waves, the result is different wavelengths for these waves (λ will be longer in the faster medium).

Reflection without Transmission

In order to discuss standing waves, we need to completely confine the wave between two endpoints – no energy can be allowed to escape via transmission. We can make such a confinement simply by cutting off the medium at the endpoints. The wave will reflect off this sudden absence of medium, and all of the energy of the incoming wave returns in the reflected wave. But does the wave flip over or stay upright in such a case?

In fact both of these results are possible, because the edge of the medium can react in one of two ways. If the edge of the medium is held fixed (i.e. not allowed to exhibit the displacement that the wave provides every other point in the medium), then the reflected wave flips over. If the edge of the medium is free to displace, then the reflected wave does not flip over.

Figure 1.5.4 – Reflection off a Fixed End

Figure 1.5.5 – Reflection off a Free End

Explaining this result is quite tricky from a perspective of forces on the end of string, and even after figuring that out, it's hard to extend it to other types of waves (this phenomenon applies to all waves, though sometimes determining what is meant by "fixed" and "free" can be tricky). But there is a nice way to use an imaginary model to achieve this result. It goes like this:

Suppose we model a single wave hitting the end of the medium with *two* waves, moving in opposite directions through the point that is the end of the medium and passing each other. Clearly the second wave doesn't exist, since there is no medium beyond the end, but its emergence from the passing point is seen as the "reflected wave," while the other wave vanishes past the passing point.

Figure 1.5.6a – Reflection Conditions Explained with Superposed Opposing Waves (Fixed End)

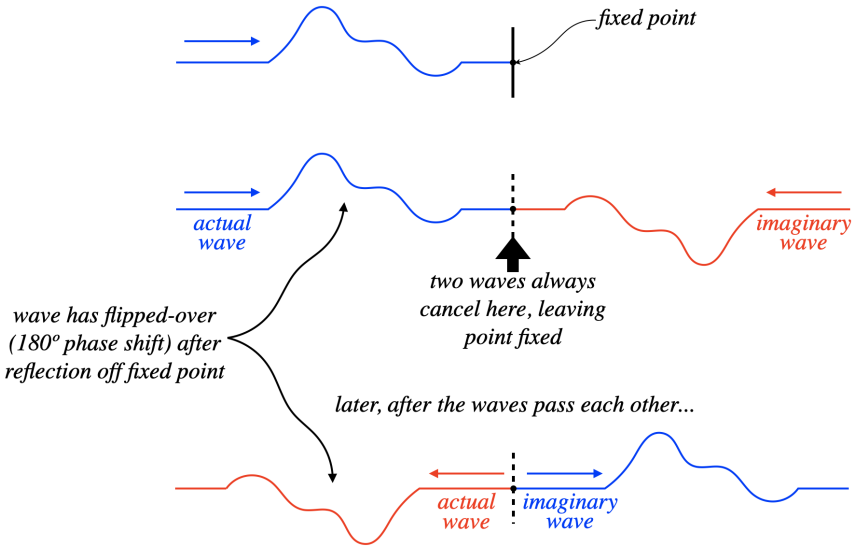
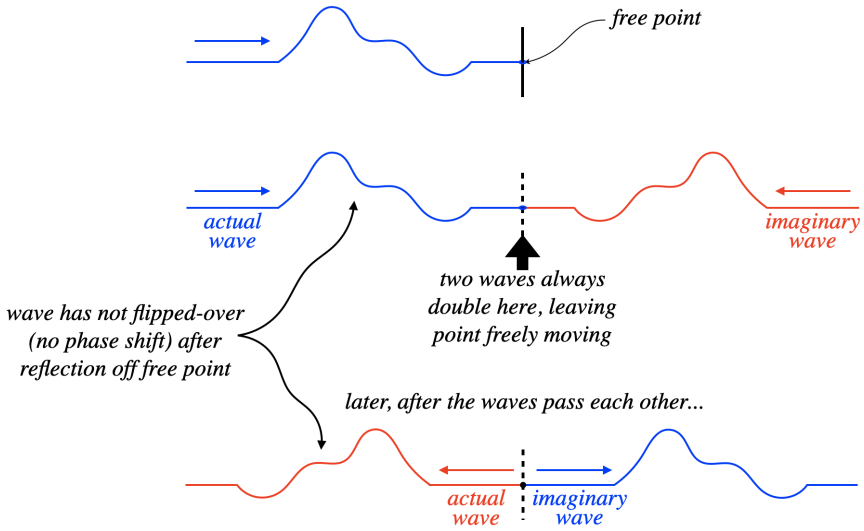


Figure 1.5.6b – Reflection Conditions Explained with Superposed Opposing Waves (Free End)



With this model, we first see that they must have the same basic waveform, and that the leading edge of one wave must correspond to the leading edge of the other. But now we ask how the imaginary wave (i.e. the second wave, before it emerges as the reflected wave) must be oriented for the passing point to be fixed or free. For the passing point to remain stationary, the superposition of the two waves at that point must result in total destructive interference. This can only happen if the second wave is inverted compared to the first wave, so when it emerges as the reflected wave, it has been flipped over. If the passing point moves freely, the two waves cannot interfere destructively, so the second wave emerges upright. Note that this analysis tells us that the free end displaces an amount equal to *twice* the amplitude, since the waves are identical and the interference is constructive.

As often as we use harmonic waves, it is useful to put the phenomenon of reflected waves in that context. When we flip over a sine or cosine wave, the result is identical to shifting that wave by a phase of π:

$$flip\ wave\ function: \quad A\cos\left(\frac{2\pi}{\lambda}x \pm \frac{2\pi}{T}t + \phi\right) \rightarrow -A\cos\left(\frac{2\pi}{\lambda}x \pm \frac{2\pi}{T}t + \phi\right) = A\cos\left(\frac{2\pi}{\lambda}x \pm \frac{2\pi}{T}t + \phi + \pi\right) \tag{1.5.1}$$

The inversion of a reflected wave after coming off a fixed end or a slower medium is therefore often referred to as a *phase shift of π*.

Standing Wave Mathematics

Now we know that we can get a wave to bounce back-and-forth between two ends of a medium, and the waves going each way are identical. If conditions for these waves are just right, their superposition results in a standing wave.

Figure 1.5.7 – Standing Wave

Let's see how this result occurs mathematically. This requires superposing two wave functions with the wave wavelength (wave number) and period (angular frequency) that are moving in opposite directions:

$$\begin{aligned} &\text{right-moving wave:} && f_1(x,t) = A\cos(kx - \omega t + \phi_1) \\ &\text{left-moving wave:} && f_2(x,t) = A\cos(kx + \omega t + \phi_2) \end{aligned} \tag{1.5.2}$$

Recall that this standing wave occurs because a single wave is bouncing back-and-forth between endpoints in the medium. The endpoints must either each be free (no phase shift) or fixed (π phase shift). For the sake of getting an easy-to-read result, we'll assume that a fixed endpoint lies at position $x = 0$. Because we are talking about a position where the wave reflects, and because the point is fixed, the two waves must be out of phase by π radians. Mathematically this means that the difference in their phase constants is π:

$$\phi_2 - \phi_1 = \pi \tag{1.5.3}$$

Plugging $x = 0$ and $\phi_2 = \phi_1 + \pi$ into the superposition of the two waves and setting the result equal to zero (that point remains fixed by our simplifying assumption), we get:

$$0 = f_{tot}(0,t) = A\cos(0 - \omega t + \phi_1) + A\cos(0 + \omega t + \phi_1 + \pi) = A\cos(-\omega t + \phi_1) - A\cos(\omega t + \phi_1) \quad \Rightarrow \quad \cos(-\omega t + \phi_1) = \cos(\omega t + \phi_1) \tag{1.5.4}$$

We can solve this for ϕ_1, which comes out to be: $0, \pm 2\pi \pm 4\pi \ldots$. We'll take the simplest solution of zero, which leaves us with the following wave function:

$$f_{tot}(x,t) = A\cos(kx-\omega t) - A\cos(kx+\omega t) \tag{1.5.5}$$

We can now apply the following trigonometric identity to get a simplified form of the standing wave function:

$$\cos(X-Y)-\cos(X+Y) = 2\sin X \sin Y \quad \Rightarrow \quad f_{SW}(x,t) = 2A\sin kx \sin\omega t = 2A\sin\left(\frac{2\pi x}{\lambda}\right)\sin\left(\frac{2\pi t}{T}\right) \tag{1.5.6}$$

All harmonic waves are collections of harmonically-oscillating points in a medium, that vary in total phase from one position to the next. Traveling waves satisfy this, but the amplitudes of these oscillators are all the same in this case. When interference occurs, the *amplitude* can vary from one position to the next as well (e.g. positions of destructive interference have zero amplitude, and positions of constructive interference have very large amplitudes), and this is evident in this result for the interference pattern we call a standing wave. This formula can be written as a collection of harmonic oscillators all with the same a period (and therefore the same sine function), but with different amplitudes at different positions:

$$f_{SW}(x,t) = \left[\mathcal{A}(x)\right]\sin\left(\frac{2\pi t}{T}\right), \qquad \mathcal{A}(x) \equiv 2A\sin\left(\frac{2\pi x}{\lambda}\right) \tag{1.5.7}$$

It is not hard to visualize this wave – it is a sine function along the x-axis, which remains in place ("standing") as its displacement at various positions oscillates with time. That is, it is exactly like the standing wave depicted in Figure 1.5.6, with the left end being the origin. There are several things to note about this standing wave:

- There are fixed points that occur at specific positions on the standing wave (when the sine function of position vanishes), called *nodes*. These are separated by a distance equal to one half the wavelength of the traveling waves. We will say that the "wavelength" of the standing wave equals the wavelength of the traveling waves that are forming it.
- The maximum displacement of the standing wave only occurs at specific positions, called *antinodes*, which are also separated by a distance of one half wavelength.
- The maximum displacement of the standing wave ($2A$) occurs when the sine functions of time and position both equal 1, and it is twice the amplitude of the traveling waves that compose it. This is referred to as the "amplitude" of the standing wave.
- The period of oscillation of the standing wave (the time it takes to get back to where it started) is the same as the period of the traveling waves that compose it (T).
- This one-dimensional function cannot be written in the form $f(x\pm vt)$, but it *is* a solution of the wave equation. The reason is that the ambiguity of the sign of $\pm v$ is washed away in the square of v in the wave equation. We originally described a wave as a phenomenon that transports energy from one position to another, and a standing wave clearly does not do this, so it is probably better described as a special time-varying interference pattern.

Note that we could have insisted that the end of the medium at the origin is free rather than fixed. This would result in no phase shift for the reflected wave, and it is left as an exercise to show that this results in a standing wave function with two cosine functions replacing the two sine functions in Equation 1.5.7.

The astute reader will notice that the standing wave equation Equation 1.5.7 has the form of a product of two functions, one of x and the other or t. and that this is precisely the form that a separable solution to the wave equation takes, as we saw at the end of Section 1.2. Indeed, we could have used the result of separation of variables instead of opposite-moving waves and trig identities, by simply applying boundary conditions. If we start from the general form of a separable wave and require that the wave remain zero at the position $x=0$ for all times, we get:

$$\mathcal{X}(0) = 0 = a\cos k(0) + b\sin k(0) \quad \Rightarrow \quad a = 0 \tag{1.5.8}$$

And if we decide that we will start our clock ($t=0$) when the wave is flat (equal to zero everywhere, when all the crests of the traveling waves moving in one direction align with all the troughs of the waves moving in the opposite direction), then:

$$\mathcal{T}(0) = 0 = c\cos\omega(0) + d\sin\omega(0) \quad \Rightarrow \quad c = 0 \tag{1.5.9}$$

And reconstructing the full wave function gives:

$$f_{SW}(x,t) = \mathcal{X}(x)\mathcal{T}(t) = bd\sin kx\sin\omega t \tag{1.5.10}$$

This is precisely Equation 1.5.7, with the amplitude $2A = bd$ and the usual identifications of the wave number k and angular frequency ω in terms of the wavelength and period.

Standing Wave Harmonics

The formula for a standing wave is still rather abstract, in that it really only restricts the behavior of the standing wave at a single point (the origin), and assumes that we know the wavelength and period. Here we will consider a different restriction, one that is more useful for physics applications. We will define a distance between two endpoints, and insist that a standing wave forms between them. We also need to

specify if the ends are held fixed or are free. If an end is fixed, it must be a node of the standing wave, and if it is free, it must be an antinode, so this greatly restricts the standing waves that can be formed. In particular, it puts very specific restrictions on the possible wavelengths a standing wave can have.

Let's start with the longest possible wavelength that a standing wave can have if its two ends are separated by a distance L. There are three possibilities in terms of the node/antinode endpoints: Both ends can be fixed (nodes), both ends can be free (antinodes), or there can be one of each type at the two ends. Note that in the first two cases, the distance between the two ends must equal one-half wavelength, while in the third case the distance between the ends is one-quarter wavelength (again, we are looking specifically at the *longest possible wavelengths* to satisfy these conditions).

Figure 1.5.8 – Longest Wavelength Standing Wave – Both Ends Fixed

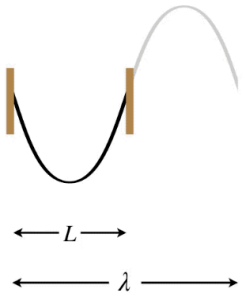


Figure 1.5.8 – Longest Wavelength Standing Wave – Both Ends Free

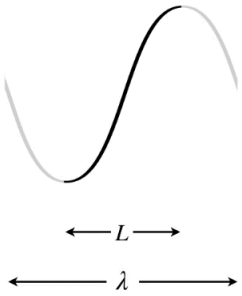
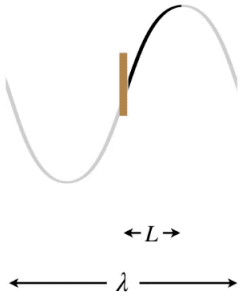


Figure 1.5.9 – Longest Wavelength Standing Wave – One End Free, One End Fixed



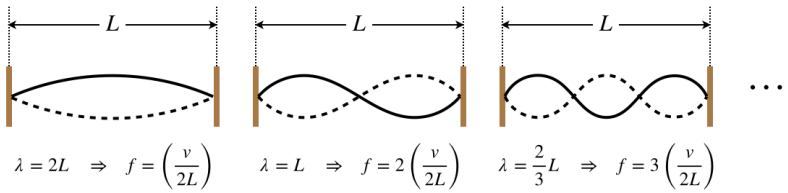
In the Figures above, the dark curves indicate the extent of the medium (i.e. that which is actually vibrating). The gray portions are only added to show the actual wavelength λ of the standing wave and how it relates to the length L of the medium.

These are not the only standing waves possible for the given length L. An infinitude of additional standing waves are possible with shorter wavelengths as well, but *only certain wavelengths will work*. We can characterize these by the number of nodes or antinodes present. The set of standing waves allowed for a given length of medium are called the *harmonics* of the system. The harmonic with the longest possible wavelength is called the *fundamental harmonic*, and the rest are numbered up from there according to frequency.

Speaking of frequency, it must be noted that the frequency of oscillation of a standing wave changes from one harmonic to the next. As we have already seen, the wavelength of the standing wave equals the wavelength of the the two opposite-moving traveling waves, and the period (or frequency) of the standing wave matches the traveling wave periods as well. If we consider a shorter-wavelength standing wave (one with more antinodes), then the wavelength of the traveling waves that make it must also be shorter. But the medium is unchanged, so the speed of those traveling waves must remain the same. This can only be true if the frequency of the traveling wave has gone up, which means the frequency of the standing wave must also go up.

We define the n^{th} harmonic as that harmonic with a frequency that is n times as great as the fundamental harmonic. Let's see what that means for the three possible endpoint conditions. We'll start with both ends fixed. For the fundamental harmonic, we found that the wavelength was double the length of the medium. The next shortest wavelength would include a single node between the two endpoints, and as allowable standing waves get shorter and shorter, we simply keep adding nodes, one at a time (this can be described as fitting an additional half-wavelength between the endpoints each time).

Figure 1.5.10 – Harmonics, Both Ends Fixed



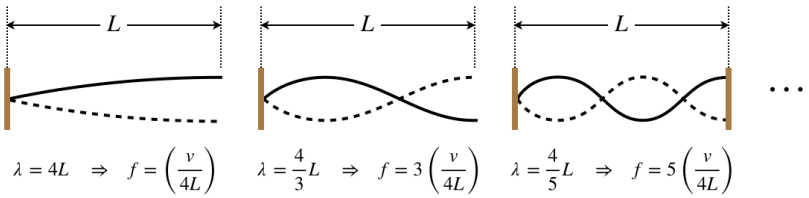
The pattern for this case is clear: The n^{th} possible standing wave has a frequency of n times the fundamental harmonic, which means that the each time we add an antinode, we get the next-highest harmonic, and the number of antinodes equals the order of the harmonic. Mathematically we summarize it this way (v is the speed of the traveling wave on the string):

$$\lambda_n = \frac{2L}{n} \quad \Rightarrow \quad f_n = n\left(\frac{v}{2L}\right), \qquad n = 1,\ 2,\ 3,\ \ldots \tag{1.5.11}$$

If we look at both ends free, we find that the same pattern emerges, which should be clear from the fact that the wavelength of the fundamental harmonic is the same when both ends are free or fixed. The only difference between the two cases are that we count the number of *nodes* to get the harmonic in the both ends free case, not antinodes, as we did for the case of both ends fixed.

When only one end is free, we get a different result when it comes to counting harmonics. We still squeeze additional half wavelength between the endpoints for the next possible wave, but the frequencies of the harmonics have a different relationship to the fundamental.

Figure 1.5.11 – Harmonics, One End Fixed, One End Free

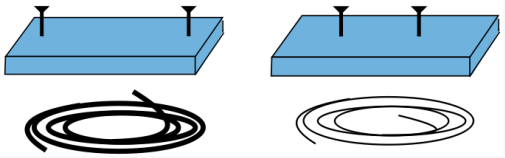


Notice that in this case each time a half-wavelength is added, the frequency jumps an amount equal to *two* fundamental harmonics. So for the case of one end fixed and the other end free, the allowed standing waves include no even-numbered harmonics. Mathematically:

$$\lambda_n = \frac{4L}{n} \quad \Rightarrow \quad f_n = n\left(\frac{v}{4L}\right), \qquad n = 1,\ 3,\ 5,\ \ldots \tag{1.5.12}$$

Example 1.5.1

Two boards with nails separated by different distances are combined with uniform strings that have different lengths and masses, to form one-string guitars.



Show that these guitars make tones of the same pitch (as determined by their fundamental harmonics) when the following quantity α is the same for both:

$$\alpha = \frac{FL}{md^2},$$

where F is the tension in the string, L is the length of the string, m is the mass of the string, and d is the distance separating the nails.

Solution

The frequencies of the fundamental harmonics must be equal, which means:

$$f_A = f_B \quad \Rightarrow \quad \frac{v_A}{\lambda_A} = \frac{v_B}{\lambda_B}$$

With both strings exhibiting their fundamental harmonics, they both have the same relationship between their wavelengths at the nail separations – in both cases the nail separation is half a wavelength:

$$\lambda_A = 2d_A\ , \quad \lambda_B = 2d_B \quad \Rightarrow \quad \frac{v_A}{d_A} = \frac{v_B}{d_B}$$

The speed of the traveling waves that create the standing wave is determined by the tension and the string density. The string is uniform, so its density is the ratio of the string's mass and its length. Therefore:

$$\mu = \frac{m}{L} \quad \Rightarrow \quad v = \sqrt{\frac{F}{\mu}} = \sqrt{\frac{FL}{m}}$$

Plugging this in above gives:

$$\frac{1}{d_A}\sqrt{\frac{F_A L_A}{m_A}} = \frac{1}{d_B}\sqrt{\frac{F_B L_B}{m_B}} \quad \Rightarrow \quad \frac{F_A L_A}{m_A d_A^2} = \frac{F_B L_B}{m_B d_B^2} = \alpha$$

Example 1.5.2

Since the time of ancient Rome, commanders of armies have known that it is prudent to have the troops break stride in their march when crossing a wooden bridge. This is because if the troops march in a synchronized cadence, they produce a periodic coordinated jolt to the bridge, which could excite one of its natural harmonic frequencies, causing a standing wave to develop in the bridge.

a. *If a marching army does create a standing wave in the bridge, what aspect of this standing wave (A, f, T, λ) would be directly responsible for causing the bridge to break apart? Explain.*
b. *Suppose a platoon comes upon a wood & rope bridge that is supported only at its two ends. The commander stops the company short of the bridge and shakes the nearest end of the bridge, testing to see if it seems strong enough to hold the troops. The bridge ripples all the way down its length, with the pulse reflecting off the other side and returning, for a round-trip time of about $2.5s$. Find the frequency of the fundamental harmonic standing wave for this bridge.*
c. *The commander decides the bridge is sturdy, and makes the tragic decision to order the company to march on. Their marching pace and spacing is such that a standing wave forms in the bridge, and the ropes break when the center of the bridge dips well below its usual point as two outer parts of the bridge surge upward. Find the marching pace of the company in steps per second.*

Solution

a. The bridge breaks apart when various components are stretched and separated so far that they can no longer hold together. This deformation of the bridge is a direct result of the amplitude of the standing wave. Put another way, the violence with which the bridge shakes is a measure of the energy put into it, and the energy in the standing wave is a function of its amplitude.

b. Call the length of the bridge L and the speed of the wave v. The time it takes the wave to travel two lengths of the bridge is given as 2.5s, and in terms of the distance traveled and speed of the wave, we have:

$$2.5s = t = \frac{2L}{v}$$

For the fundamental harmonic, the length of the bridge (which is fixed at both ends) is one-half wavelength, so plugging in a half wavelength for L gives:

$$2.5s = \frac{2\,(0.5\lambda)}{v} = \frac{\lambda}{v} = \frac{1}{f} \quad \Rightarrow \quad f = 0.40Hz$$

c. The description of the standing wave makes it clear that it has three antinodes, which means it is the 3rd harmonic. The two ends of the wave are fixed, so the third harmonic occurs at three times the fundamental frequency, or $1.2Hz$ For the footfalls to excite this harmonic, they need to match this frequency, so the marching pace is 1.2 steps per second.

Alert

If you are a musician, you likely have heard of overtones. At the simplest level (like one-dimensional standing waves with both ends fixed or free), these are synonymous with harmonics. But in the one-dimensional case when one end is free, or in the case of two-dimensional

standing waves (like those produced by a membrane on a drum), these definitions diverge. We will not go into the details of these divergences, and the rare times we refer to the "first overtone," we will mean simply the next highest allowable harmonic.

Energy In Standing Waves

Let's consider the case of a second harmonic standing wave on a string between two fixed ends. We know the following things to be true:

- Between the endpoints, there is exactly one full wave moving right and an identical full wave moving left at all times.
- Each of these waves contains an amount of energy that is proportional to the square of its amplitude.
- The standing wave has an amplitude twice as great as the amplitude of each individual traveling wave.

So the question is, doesn't doubling the amplitude mean the standing wave has *four times the energy* of an individual traveling wave? If it does mean this, where does this extra energy come from, if there are only two such waves providing energy?

This apparent paradox stems from something we discussed earlier – *it is dangerous to think of a standing wave in the same context as a traveling wave*! This is especially true in the context of energy distribution. Let's consider the energy of a single particle in a medium as a harmonic wave passes through. Such a particle is following harmonic motion, so if it happens to be at the crest or the trough of the wave, then its kinetic energy is zero, while its potential energy is a maximum. Conversely, if it is at the middle, then it has its maximum kinetic energy and no potential energy. *But no matter where it is in the phase of the wave, its energy is the same.*

Now compare that with a particle in the medium of a standing wave. If the particle is at a node, then it never moves, and is never displaced from equilibrium, so its energy is zero. A particle at an antinode, on the other hand, has lots of energy. The amplitude of its harmonic motion is twice the amplitude that a particle on one of the two traveling waves would have, if the second wave wasn't there.

The bottom line is that the standing wave, *when viewed as an interference pattern*, clearly just redistributes the energy of the two traveling waves (which themselves distribute the energy uniformly), taking energy away from some regions of the medium and giving it to others. With some clever calculus, we can show that this works out exactly.

If the string has a linear density of μ, then an infinitesimal segment of the string of length dx has a tiny mass of $dm = \mu dx$. A traveling wave has every such infinitesimal segment oscillating with the same amplitude, so every particle on the string contributes the same infinitesimal energy, and adding these contributions for a full wavelength gives:

$$E_{traveling\ wave} = \int_0^\lambda \frac{1}{2} dm\ \omega^2 A^2 = \int_0^\lambda \frac{1}{2} \mu dx\ \omega^2 A^2 \tag{1.5.13}$$

The density of the string, the frequency of oscillation, and the amplitude of oscillation are the same for every particle in the string, so they do not vary with x, which makes the integral simple to perform:

$$E_{traveling\ wave} = \frac{1}{2} \mu \lambda \omega^2 A^2 \tag{1.5.14}$$

The segments of the string for a standing wave behave differently. They all vibrate harmonically (with the nodes exhibiting zero vibration), but they reach different maximum displacements. Put another way, a standing wave is a collection of an infinite number of harmonic oscillators, all with different amplitudes. So we need to write down the energy for each particle, and add them all up. The waveform of the standing wave gives us the amplitude (which we will call $a(x)$) of particle oscillation as a function of position x, so from Equation 1.5.7, we have:

$$Amplitude\ of\ medium\ at\ x = a(x) = 2A \sin\left(\frac{2\pi x}{\lambda}\right) \tag{1.5.15}$$

Recall that A is the amplitude of the two traveling waves that are interfering. The energy of this tiny piece of the string is:

$$dE = \frac{1}{2} dm\ \omega^2 \left[a(x)\right]^2 \tag{1.5.16}$$

Putting in $dm = \mu dx$ and $a(x)$ and integrating over the full wavelength of the wave, we get:

$$E_{standing\ wave} = \int_0^\lambda \frac{1}{2} \mu dx\ \omega^2 \left[2A \sin\left(\frac{2\pi x}{\lambda}\right)\right]^2 \tag{1.5.17}$$

Making the substitution $u \equiv \frac{2\pi x}{\lambda}$ leaves an integral that is easy to look up, and gives the following answer:

$$E_{standing\ wave} = \frac{\mu\lambda\omega^2 A^2}{\pi}\int_0^{2\pi} \sin^2 u\ du = \mu\lambda\omega^2 A^2 \tag{1.5.18}$$

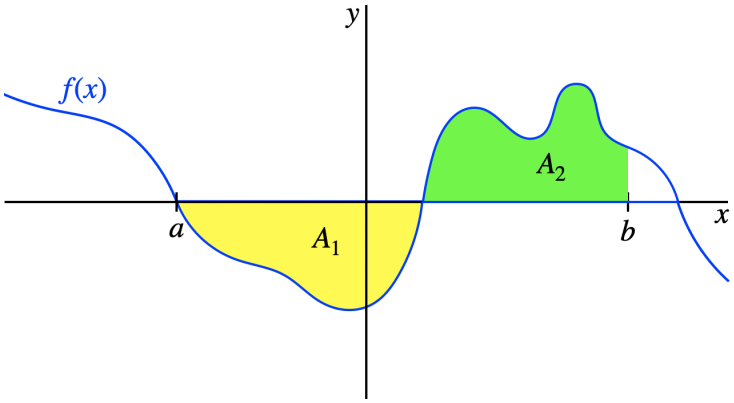
Comparing this with the answer for the traveling wave, we see that it is twice as much – the energy content of the standing wave equals the sum of the energies of the two traveling waves that interfere to create it. If we want to write the energy contained in a single wavelength of a standing wave in terms of the standing wave's "amplitude" (the amplitude of the harmonic motion located at an antinode), we have:

$$\mathcal{A} = 2A \quad \rightarrow \quad E_{standing\ wave} = \frac{1}{4}\mu\lambda\omega^2\mathcal{A}^2 \tag{1.5.19}$$

1.6: Some Important Math Tricks

Odd and Even Functions

We are well familiar with cases where integrals of functions can equal zero. Essentially all that is required is that the function creates an equal amount of area above the horizontal axis as below it:

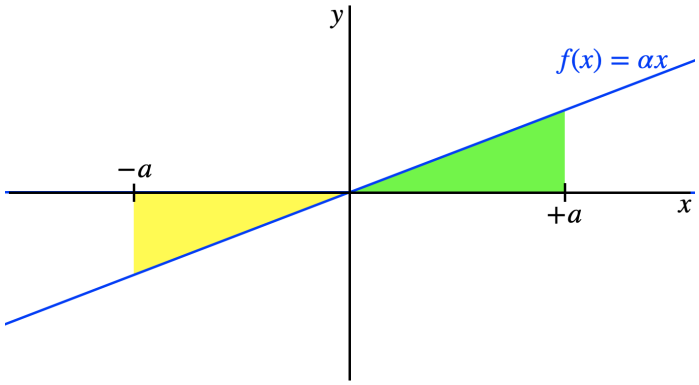


If the areas A_1 and $A2$ are equal, then:

$$\int_a^b f(x)\,dx = 0 \tag{1.6.1}$$

This fact allows us to solve certain integrals very fast, if we know something about the symmetry of the function we are integrating over the interval of integration. For simplicity, we will limit this discussion to symmetries about the vertical axis, but keep in mind that these can be shifted in either direction by a simple change of x-coordinates. The simplest function that exhibits this is a line that passes through the origin, integrated between limits equidistant on both sides of the y-axis:

Figure 1.6.2 – Line Through the Origin



This integral obviously vanishes thanks to similar triangles, but we can also confirm it "the long way":

$$\int_{-a}^{+a} f(x)\,dx = \int_{-a}^{+a} \alpha x dx = \left[\frac{1}{2}\alpha x^2\right]_{-a}^{+a} = 0 \tag{1.6.2}$$

It should be clear that we should get this result for functions like $f(x) = \alpha x^3$, $f(x) = \alpha x^1 1$, and indeed any function that is just an odd power of x, because the integral will result in a function that has an even power, and the difference of the two endpoints will always vanish:

$$\int_{-a}^{+a} f(x)\,dx = \int_{-a}^{+a} \alpha x^n dx = \left[\frac{1}{n+1}\alpha x^{(n+1)}\right]_{-a}^{+a} = 0\,, \quad n \text{ odd} \tag{1.6.3}$$

It should be equally clear that adding two such functions together results in another function with the same property, since the integral of each term in the sum vanishes. Taking this to its extreme, it means that any function that can even be expressed as a power series that

includes only odd-powered terms will also have a vanishing integral over limits equidistant from the $x = 0$ axis. Such functions are called *odd functions*, for obvious reasons.

The counterpart of these functions are *even functions*, which are expressible as a power series of even powers of the argument. [Functions that are odd are said to have *odd parity*, and even functions *even parity*. Functions that fall into neither category are said to not have *definite parity*.] Even functions are similar to odd functions in that the areas they define with the horizontal axis are equal on both sides of the vertical axis, but the difference is that these areas have the *same sign*, so they don't cancel. Knowing a function is even does help simplify the work a bit (not as much as just knowing the integral is zero, of course!), in that we can change one of the limits of integration to zero, and multiply by two:

$$\int_{-a}^{+a} f_{even}(x)\,dx = 2\int_{0}^{+a} f_{even}(x)\,dx \tag{1.6.4}$$

The most common (and in our case, most useful) examples of odd and even functions are sine and cosine, respectively:

$$\sin x = \frac{1}{1!}x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \ldots \qquad \cos x = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \ldots \tag{1.6.5}$$

The reader can verify for themself that the integral of these functions over an interval symmetric about the x-axis gives the expected results for odd and even functions. As stated above, this odd/even property doesn't only apply across the origin – a function can be odd or even over any specific interval. For example, a sine function is odd over the interval from 0 to 2π, but even over the interval from 0 to π. The cosine function is precisely the opposite – even over the interval from 0 to 2π, and odd over the interval from 0 to π.

There is one more property of these kinds of functions that is important to point out. Whenever a new function is formed from the product of two odd or two even functions, the result is an even function. To see this, consider multiplying all the terms in the power series. The powers add in each product, and adding two odd or two even numbers results in an even number. This should make it equally clear that the product of one odd function and one even function results in an odd function. So the integral of a product of two functions may look very complicated, but if one of the functions is odd and the other is even, and the limits of the integral are symmetrically-placed across the vertical axis, then we know immediately that the integral vanishes. If there are more than two functions multiplying each other, then the parity of any pair can first be determined, then the parity of that pair can be combined with the third function's parity, and so on.

Orthogonal Functions

While the details are slightly beyond the scope of this course, it's useful to know that the properties of, and operations involving, vectors that we first encountered in Physics 9HA extend far beyond entities that "have magnitude and direction". For example, similar properties can be attributed to polynomials and functions in general (which are expressible as power series – polynomials with an infinite number of terms). In particular, we can create a consistent definition of the "orthogonality" of two functions. We define two functions to be *orthogonal* when the integral of their product over all values vanishes (sometimes we limit the range of the integral, like when the functions are periodic and the integral is clearly repeating itself).

You can think of two functions as vectors, and the integral of their product as their dot product. This perspective is quite useful in many contexts, as well as being strictly accurate in a mathematical sense, though it is more abstract than what we have seen so far. So clearly all odd functions are orthogonal with all even functions. But having opposite parity across the origin is not the only mode by which two functions can be orthogonal. We will see examples of this as we go through this course (as well as in future physics classes), but right now the most important example involves harmonic functions. We already know that $\sin k_1 x$ is orthogonal to $\cos k_2 x$ for any values of k_1 and k_2, over any interval that is symmetric about the origin, thanks to their parities. But there is another example involving a pair of sine functions or a pair of cosine functions, though the arguments of these functions and the intervals of integration are restricted.

$$\int_0^{2\pi} \cos(m\theta)\cos(n\theta)d\theta = \int_0^{2\pi} \sin(m\theta)\sin(n\theta)d\theta = \begin{cases} \pi & m = n \\ 0 & m \neq n \end{cases} \qquad m,\ n \text{ are integers} \tag{1.6.6}$$

[Actually, the limits of integration do not need to be 0 to 2π for this result to hold. Any limits that *differ* by 2π will produce the same result. That is, the integral just has to be over a single full cycle, starting at any phase.]

This remarkable fact indicates that these cosine and sign functions are orthogonal to each other over this interval of integration whenever the integers m and n are not equal. And we already know that the cosine and sine functions are orthogonal to each other (even when $m = n$) over this interval, thanks to their parities. This is another way in which the view of these functions as vectors differs from what we are used to: There are an infinite number of these vectors, all perpendicular to each other!

A Brief Foray Into Abstract Mathematics

Let's take a moment to have a closer look at this notion of treating functions like vectors. While we will use some of the notation that follows only sparingly in the chapters to come, in more advanced courses it becomes the standard, so it is a good idea to get exposed to it early.

We saw in our study of 4-vectors in Physics 9HB that the vector itself is a well-defined object, independent of the coordinate system we use to describe it (and define its components). We know of a similar concept for functions – changing variables. We can write out the function $f(x)$ or we can make the substitution $y = \alpha x + 3$ and write out the function in terms of y. In some abstract (and technically imprecise) sense, we can think of f as the "vector" and $f(x)$ as the components of that vector. One of the things that makes this description difficult to compare with vectors we are used to is that these "function vectors" have an infinite number of components – one for each value of x. But the notion of changing variables to get a whole new (infinite) set of components for the same vector is a reasonable one.

There is a notation that has been invented, by a physicist and in the context of quantum theory, that does a good job of expressing this functions-as-vectors idea. It is called *Dirac bra-ket notation*. There is much more to this notation than will be covered here (most notably the role of complex numbers), but the basics are as follows:

- A bracket "$\langle\ |\ \rangle$" is broken into two halves, the left half "$\langle\ |$" known as a "bra", and the right half "$|\ \rangle$" called a "ket".
- Whether a bra or a ket, it is a vector, and the combined bracket is the dot product between the two vectors represented by the bra on the left and the key on the right:

$$\langle u| \leftrightarrow \vec{u}\,, \quad |v\rangle \leftrightarrow \vec{v} \quad \Rightarrow \quad \langle u|v\rangle \leftrightarrow \vec{u}\cdot\vec{v} \tag{1.6.7}$$

- We consider the function to be an abstract vector $|f\rangle$, and the variable used to in the function as a sort of unit vector $\langle x|$. Taking the dot product of a vector with a unit vector yield the component of the vector along that unit vector's direction:

$$\hat{i}\cdot\vec{v} = v_x \quad \leftrightarrow \quad \langle x|f\rangle = f(x) \tag{1.6.8}$$

- The dot product between two vectors can be written in terms of the sum of the product of their components in the same coordinate system. In the case of functions, there are a continuum of unit vectors, so the sum taken of components is an integral:

$$\vec{u}\cdot\vec{v} = \left(\hat{i}\cdot\vec{u}\right)\left(\hat{i}\cdot\vec{v}\right) + \left(\hat{j}\cdot\vec{u}\right)\left(\hat{j}\cdot\vec{v}\right) + \left(\hat{k}\cdot\vec{u}\right)\left(\hat{k}\cdot\vec{v}\right) = u_x v_x + u_y v_y + u_z v_z \quad \leftrightarrow \quad \langle f|g\rangle = \int \langle f|x\rangle\,\langle x|g\rangle\,dx = \int f(x)\,g(x)\,dx \tag{1.6.9}$$

While the placements on the left or right side of the brackets shown above are technically important, this is intended as a basic introduction to this notation and the notion of functions as vectors, and not a formal exposition. The official name for this functions-as-vectors formalism is *Hilbert space*.

1.7: Fourier Analysis

Fourier Series

Continuing with the idea in the previous section that functions can be treated as vectors, we note that we can construct an arbitrary vector from a linear combination of the full set of orthogonal unit vectors. It turns out that we can construct a certain family of functions from a linear combination of harmonically-linked cosines and sines (which we saw in the previous section are mutually orthogonal). Of course, as there are an infinite number of these "orthogonal" vectors, the linear combination could be an infinite sum. The family of functions that can be constructed in this way happen to be those that are periodic. We will assert without formal proof that *any* periodic function can be written as a linear combination of cosines and sines. Showing the specific form of this linear combination requires rewriting Equation 1.6.6 in a more general form – one that accommodates periodic functions of arbitrary wavelengths. Changing the angle θ, measured, in radians into a form that depends upon position in space gives:

$$\theta = \frac{2\pi}{\lambda}x \tag{1.7.1}$$

The limits of integration in Equation 1.6.6 are for the variable θ, so for x these limits are $\pm\frac{\lambda}{2}$. With these alterations, Equation 1.6.6 becomes:

$$\int_0^{\lambda} \cos\left(\frac{2\pi m}{\lambda}x\right)\cos\left(\frac{2\pi n}{\lambda}x\right)dx = \int_0^{\lambda} \sin\left(\frac{2\pi m}{\lambda}x\right)\sin\left(\frac{2\pi n}{\lambda}x\right)dx = \begin{cases} \frac{\lambda}{2} & m = n \\ 0 & m \neq n \end{cases} \qquad m,\ n \text{ are integers} \tag{1.7.2}$$

The linear combination of this infinite set of orthogonal functions can then construct any function that is periodic with a wavelength equal to λ. Defining the coefficients of the cosines and sines separately, and summing over all integers, we have:

$$f(x) = \sum_{n=-\infty}^{n=\infty}\left[a_n \cos\left(\frac{2\pi n}{\lambda}x\right) + b_n \sin\left(\frac{2\pi n}{\lambda}x\right)\right] \tag{1.7.3}$$

The sum over the negative integers is redundant with the sum over the positive integers, and clearly the parities of cosine and sine require that $a_{-n} = a_n$ and $b_{-n} = b_n$, meaning we can change this sum over all the integers into twice the sum over the positive integers (plus $n = 0$). The constant b_o is pointless, since what it multiplies is identically zero. And the constant a_o multiplies 1, so we can rewrite this expansion as:

$$f(x) = a_o + \sum_{n=1}^{n=\infty}\left[2a_n \cos\left(\frac{2\pi n}{\lambda}x\right) + 2b_n \sin\left(\frac{2\pi n}{\lambda}x\right)\right] \tag{1.7.4}$$

It is customary not to carry around the factors of 2, choosing instead to redefine the constants as half of those given above. This gives us our final expression of what is called the *Fourier series expansion* of the periodic function $f(x)$:

$$f(x) = \frac{a_o}{2} + \sum_{n=1}^{n=\infty}\left[a_n \cos\left(\frac{2\pi n}{\lambda}x\right) + b_n \sin\left(\frac{2\pi n}{\lambda}x\right)\right] \tag{1.7.5}$$

One can look at this as a sum of harmonic functions with wavelengths equal to or shorter than the wavelength of the periodic wave. To be exact, the wavelength of the n^{th} harmonic wave in the sum is $\frac{\lambda}{n}$, so the wavelengths of the harmonic functions are integer fractions of the periodic wave's wavelength. Just to drive home the point that this expansion requires that the function is periodic, we can check this directly:

$$f(x+\lambda) = \frac{a_o}{2} + \sum_{n=1}^{n=\infty}\left[a_n \cos\left(\frac{2\pi n}{\lambda}(x+\lambda)\right) + b_n \sin\left(\frac{2\pi n}{\lambda}(x+\lambda)\right)\right] = \frac{a_o}{2} + \tag{1.7.6}$$
$$\sum_{n=1}^{n=\infty}\left[a_n \cos\left(\frac{2\pi n}{\lambda}x + 2\pi n\right) + b_n \sin\left(\frac{2\pi n}{\lambda}x + 2\pi n\right)\right]$$

Of course, for any integer n, $\cos(\theta + 2\pi n) = \cos\theta$ and $\sin(\theta + 2\pi n) = \sin\theta$, so this comes back to the original series, confirming that $f(x+\lambda) = f(x)$.

What good is this series if we don't have a way to figure out what the coefficients are? Well, it turns out that we do have a way! As a hint for how to do this, we can once again turn to what we know about vectors. If the cosine and sine functions are the equivalent of

unit vectors, then the a_n and b_n coefficients are the components of the vector, and we know a way to find the components of a vector. For example:

$$v_x = \hat{i} \cdot \vec{v} \tag{1.7.7}$$

Using our identification of these functions as vectors, we should be able to use an integral (the function version of a dot product) to get the constants. Well, the *reason* that the dot product with the unit vector works is that the dot product vanishes with every unit vector accept the one being used in th edot product. The orthogonal functions do the same thing! Taking this "integral dot product":

$$\int_0^\lambda f(x)\cos\left(\frac{2\pi m}{\lambda}x\right)dx = \int_0^\lambda \left(\frac{a_o}{2} + \sum_{n=1}^{n=\infty}\left[a_n\cos\left(\frac{2\pi n}{\lambda}x\right) + b_n\sin\left(\frac{2\pi n}{\lambda}x\right)\right]\right)\cos\left(\frac{2\pi m}{\lambda}x\right)dx \tag{1.7.8}$$

Every integral in this expression involves orthogonal functions *except* for the integral of the two cosines where $n = m$. Using Equation 1.7.2 gives:

$$\int_0^\lambda f(x)\cos\left(\frac{2\pi m}{\lambda}x\right)dx = a_n\int_0^\lambda \cos\left(\frac{2\pi m}{\lambda}x\right)\cos\left(\frac{2\pi n}{\lambda}x\right)dx \quad \Rightarrow \quad a_n = \frac{2}{\lambda}\int_0^\lambda f(x)\cos\left(\frac{2\pi m}{\lambda}x\right)dx \tag{1.7.9}$$

We can compute the b_n coefficients the same way:

$$b_n = \frac{2}{\lambda}\int_0^\lambda f(x)\sin\left(\frac{2\pi m}{\lambda}x\right)dx \tag{1.7.10}$$

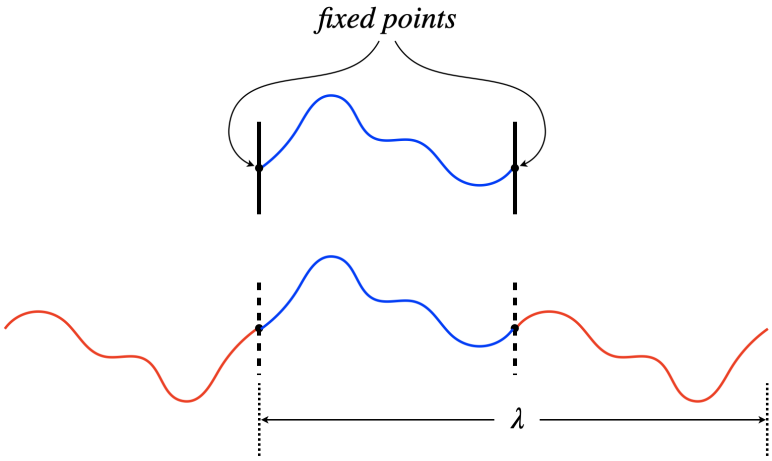
What we effectively have here is a means of writing down the "recipe" of any periodic wave as a collection of numbers, a_n and b_n. This collection may be infinite in number, or possibly not, but even if it is infinite, it may be possible in some cases to just keep the most important terms in the series as an approximation. This "recipe" is sometimes referred to as the *spectral content* of the wave. This is because it tells us the contribution of each harmonic wave that comprises it, with the harmonic waves differing from each other by an integer fraction of the wavelength of the periodic wave. In a later chapter, we will see how this concept can be generalized to non-periodic waves as well, and we will find that this requires that the spectral content take into account *all* wavelengths, not just integer fractions. That is, while the "spectrum" for the periodic wave consists of discrete choices of harmonic wavelengths, the spectrum of a non-periodic wave (like a wave pulse) requires the whole continuum of possible harmonic wavelengths.

Application to Standing Waves

It's easy to get lost in all the math above and forget what all of it means, so let's take a step back and take a look at the big picture. One place where this math becomes very useful is with standing waves. Consider a string that is fixed at both ends. We found that it can vibrate in an infinite number of harmonics, defined by the number of half-wavelengths that fit between the two fixed ends. We also found that the solutions to the wave equation that describe standing waves are separable, with the frequency of vibration uniquely defined for each harmonic. But what if the wave is not one of these nice sinusoids?

We know standing waves occur because two identical waves moving in opposite directions interfere with each other. We also know that these two opposite-moving waves are arranged so that the boundary conditions hold (in the case we are considering, these are fixed endpoints). We've seen a model for this when we first discussed wave reflection, in Figure 1.5.6a. In that case, we considered only two transient wave pulses, but with *both* ends fixed, this dance repeats over and over, as the wave bounces back and forth. So rather than just one imaginary wave pulse to interact with the actual wave, we have two infinitely-long periodic waves that continuously interact with each other, as we see in Figure 1.5.7, but with an arbitrary function rather than a sinusoid:

Figure 1.7.1 – Opposing Wave Model of a General Standing Wave



We know that the opposite-moving waves must superpose perfectly to continually result in total cancellation at the fixed endpoints, so this gives us the shape of the combination of those waves outside the endpoints. Just like the case of the wave pulse, these "imaginary" sections have to be reflections of what is between the fixed points. The amazing upshot here is that this means that the infinitely-long standing wave (extended beyond the endpoints) is periodic in space, with a wavelength equal to twice the distance separating the two endpoints.

Now that we know that a standing wave of any shape is equivalent to a half-wavelength segment of a full periodic wave (such that the boundary conditions are met at the endpoints), we can employ the power of Fourier series to standing waves. Defining the left fixed end as the origin $x = 0$, and insisting that the other end remain fixed (the ends are separated by a distance L), we find that the possible harmonic standing wave waveforms (the first three of which are depicted in Figure 1.5.10) have the form:

$$g_n(x) = A\sin\left(\frac{n\pi}{L}x\right), \tag{1.7.11}$$

where n is the number of antinodes present. We know that a full arbitrary waveform with wavelength equal to $2L$ can be expressed in terms of a Fourier series, according to Equation 1.7.5. This waveform must remain zero at all times at $x = 0$ and $x = L = \frac{\lambda}{2}$, and plugging this into the series gives us that all of the a coefficients vanish, leaving only:

$$f(x) = \sum_{n=1}^{n=\infty} b_n \sin\left(\frac{2\pi n}{\lambda}x\right) = \sum_{n=1}^{n=\infty} b_n \sin\left(\frac{\pi n}{L}x\right) \tag{1.7.12}$$

So an arbitrary standing wave waveform can be written in terms of a sum of harmonic wave functions. What is more, we can determine exactly the recipe for this sum. Noting that $\lambda = 2L$, we get:

$$b_n = \frac{2}{\lambda}\int_0^{\lambda} f(x)\sin\left(\frac{2\pi n}{\lambda}x\right)dx = \frac{1}{L}\int_0^{2L} f(x)\sin\left(\frac{\pi n}{L}x\right)dx \tag{1.7.13}$$

This can be simplified still further. The wave in the region from $x = L$ to $x = 2L$ is the same as in the region from $x = 0$ to $x = L$, except that it is reflected both horizontally and vertically. But the same is true of every sine function in the series! This means that the integral that computes b_n is the same from $x = L$ to $x = 2L$ as it is from $x = 0$ to $x = L$. This allows us to change the limits of integration from $0 \to 2L$ to $0 \to L$, if we just multiply the integral by 2. This gives:

$$b_n = \frac{2}{L}\int_0^{L} f(x)\sin\left(\frac{\pi n}{L}x\right)dx \tag{1.7.14}$$

What About Time Dependence?

We have focused entirely on the waveform of the standing wave, but it is also evolving in time. How do we handle that? We found in our application of separation of variables to standing waves that every harmonic has associated with it a unique frequency. This is also evident from the fact that the speed of every wave in the medium is the same, and $v = \lambda f$, so there is a different frequency associated with every wavelength. The total standing wave is a linear combination of all the harmonic traveling waves, each of those being separable according to Equation 1.2.18:

$$f(x,t)=\sum_{n=1}^{n=\infty} f_n(x,t)=\sum_{n=1}^{n=\infty}\left[\left(a_n\cos k_n x+b_n\sin k_n x\right)\left(c_n\cos\omega_n t+d_n\sin\omega_n t\right)\right] \tag{1.7.15}$$

If we define our $t=0$ time as the time when we see the function $f(x)$ on the string (i.e. $f(x)=f(x,0)$), and we include our boundary conditions so that the x portion of the wave is the Fourier series, then we have:

$$f(x,0)=f(x)=\sum_{n=1}^{n=\infty}\left[\left(0+b_n\sin\left(\frac{n\pi}{L}x\right)\right)\left(c_n\cos 0+d_n\sin 0\right)\right] \quad\Rightarrow\quad c_n=1 \tag{1.7.16}$$

$$\Rightarrow\quad f(x,t)=\sum_{n=1}^{n=\infty} b_n\sin\left(\frac{n\pi}{L}x\right)\left(\cos\omega_n t+d_n\sin\omega_n t\right)$$

The only free parameters remaining are the d_n's (the ω_n's are obtained from $k_n=\frac{n\pi}{L}$, as we preseumably know the speed of the wave on the string), and these can only be determined by some additional time condition. They provide the relative phases of the individual harmonic waves. That is, we know how all of the harmonic waves are aligned at the moment in time $t=0$, and we know the frequency of oscillation of each of these waves, but we don't know how their phases compare, so we need one more piece of information to know precisely how the standing wave will evolve. For example, we don't know if the standing wave at $t=0$ is stationary or moving. Quite often, we are interested in cases where the string is distorted and released from rest, and asked how the standing wave evolves. In this case, we *do* have a definitive answer. In this case, the first derivative of the function with respect to time is zero, which means:

$$0=\left[\frac{\partial}{\partial t}f\right]_{t=0}=\sum_{n=1}^{n=\infty} b_n\sin\left(\frac{n\pi}{L}x\right)\left[-\omega_n\sin\omega_n t+d_n\omega_n\cos\omega_n t\right]_{t=0} \quad\Rightarrow\quad d_n=0 \tag{1.7.17}$$

And we get the simple final result:

$$f(x,t)=\sum_{n=1}^{n=\infty} b_n\sin\left(\frac{n\pi}{L}x\right)\cos\omega_n t\,,\quad \text{where: } b_n=\frac{2}{L}\int_0^L f(x)\sin\left(\frac{\pi n}{L}x\right)dx\,,\ \text{and } \omega_n=\frac{\pi n v}{L} \tag{1.7.18}$$

To summarize, in words...

If we are given an arbitrary starting configuration of a string between two fixed points separated by L (or in general, any standing wave with known boundary conditions), we can use the Fourier series to decompose that configuration into a sum of many (perhaps an infinite number) of harmonic functions with wavelengths that are integer fractions of $2L$. We can compute the coefficients that multiply each harmonic function (the "recipe" of the starting configuration) using the overlap integral of the the starting function with each harmonic function. Then, with some information about the starting motion of the string (e.g. it is stationary), we can construct the full final solution by multiplying each harmonic function in x by a harmonic function in t with the correct frequency, where $\omega_n=k_n v$ and v is the speed of a traveling wave on that string.

CHAPTER OVERVIEW

2: Physical Optics

2.1: Light as a Wave

What is "Waving"?

The jump from mechanical waves to sound was a difficult one, mainly because the "displacement" of the wave changed from matter that oscillates back-and-forth, to (in the case of sound in a gas) oscillations in pressure or density. This difficulty gets greatly magnified for the case of light. We know that light is a wave based on how it behaves – it exhibits the same properties of other waves we have examined – it interferes with itself, it follows an inverse-square law for intensity (brightness), and so on. But we also know that we can see light from the sun, moon, and stars, which means that light waves can travel through the vacuum of space. Unlike every other wave we have seen, it doesn't require any medium at all! So what do we use as the "displacement" for our wave function?

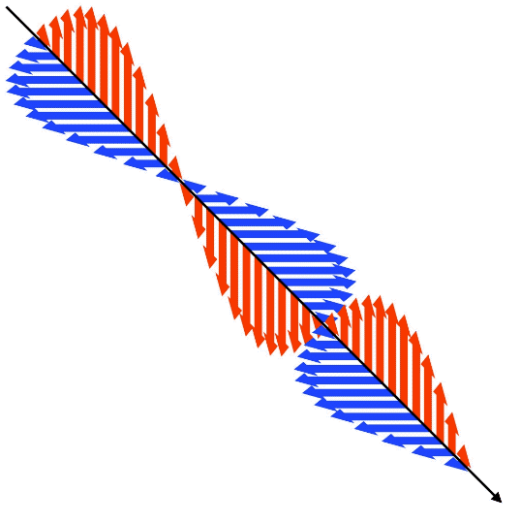
Back in the 19th century, physicists studied extensively the subjects of electricity (lightning, shocking your finger on a doorknob, balloons sticking to your hair, etc.) and magnetism (compasses, sticking things to your refrigerator, etc.). It started becoming clear that the two forces, while different, had some links. Electric currents were found to affect compass needles, and magnets moving near wires were found to create electric currents. It all came together with an amazing (for the time) effort in mathematics by a man named James Clerk Maxwell. He showed that changing electric fields could induce magnetic fields, while changing magnetic fields could in turn induce electric fields. This is a recipe for propagation of these fields, and the equation he derived for this propagation was exactly the wave equation! So he predicted, from results taken from experiments in electricity and magnetism, that an *electromagnetic wave* could be produced. The wave equation included physical constants from both electricity and magnetism, and extracting the wave speed from this equation resulted in a number Maxwell was already familiar with – the speed of light. It is traditional to denote this speed with a lower-case 'c':

$$c = 3.0 \times 10^8 \frac{m}{s} \tag{2.1.1}$$

So the "displacement" of such a wave is actually the electric and magnetic field vectors (both types of fields are waving simultaneously, with each inducing the other) in the space through which the light wave is traveling. Don't worry that this doesn't make much sense right now – it should be a bit clearer when you get to Physics 9C and study electricity & magnetism.

Okay, so for light we now have the wave speed and the "displacement." Let's address a couple other elements of light as a wave. First, a medium is not needed, as electric and magnetic field can exist in a vacuum. The presence of a medium (such as air or water) *does* effect the electric and magnetic fields, because media are made up of atoms, which are composed of positive and negative electric charges. Because of this, the speed of light within a medium is different (slower) than its speed in a vacuum. Mathematics and experiments show that light is a transverse wave – the electric and magnetic field vectors point in directions that are perpendicular to the direction of motion of the light wave (and as it turns out, they also rare always perpendicular to each other).

Figure 2.1.1 – Electromagnetic Wave



The red arrows in the figure above represent electric field vectors, and blue arrows magnetic field vectors. Specifically, this is a *plane-polarized* EM wave, which means the field vectors of a given type remain in a single plane. We will discuss plane polarization soon, but it should be noted that EM waves do not have to behave this way, so long as the electric and magnetic field vectors remain perpendicular to each other and to the direction of motion. For example, a *circularly polarized* EM wave features electric and magnetic field vectors that circulate their directions (while remaining perpendicular to each other and the direction of motion) as the wave propagates, like the hands of an analog clock, and can do so in a clockwise or counterclockwise manner.

Finally, we need to say two things about light perception. For sound, intensity (proportional to amplitude-squared) is perceived as loudness, and for light it is brightness. For sound, frequency is perceived as pitch, and for visible light it is perceived as color. The qualification "visible" must be appended because we can only see a very limits spectrum of light frequencies, the rainbow of colors often described with the acronym ROYGBIV (Red, Orange, Yellow, Green, Blue, Indigo, Violet). The red end of the visible spectrum exhibit the lowest frequencies, and the violet the highest. But of course light waves can come in frequencies much lower and much higher, and at various arbitrary cutoffs, they are given names you have probably heard before. In order of increasing frequency below the red end of the visible spectrum we have: *radio waves*, *microwaves*, and *infrared*; and above the violet end of the spectrum: *ultraviolet*, *x-rays*, and *gamma rays*.

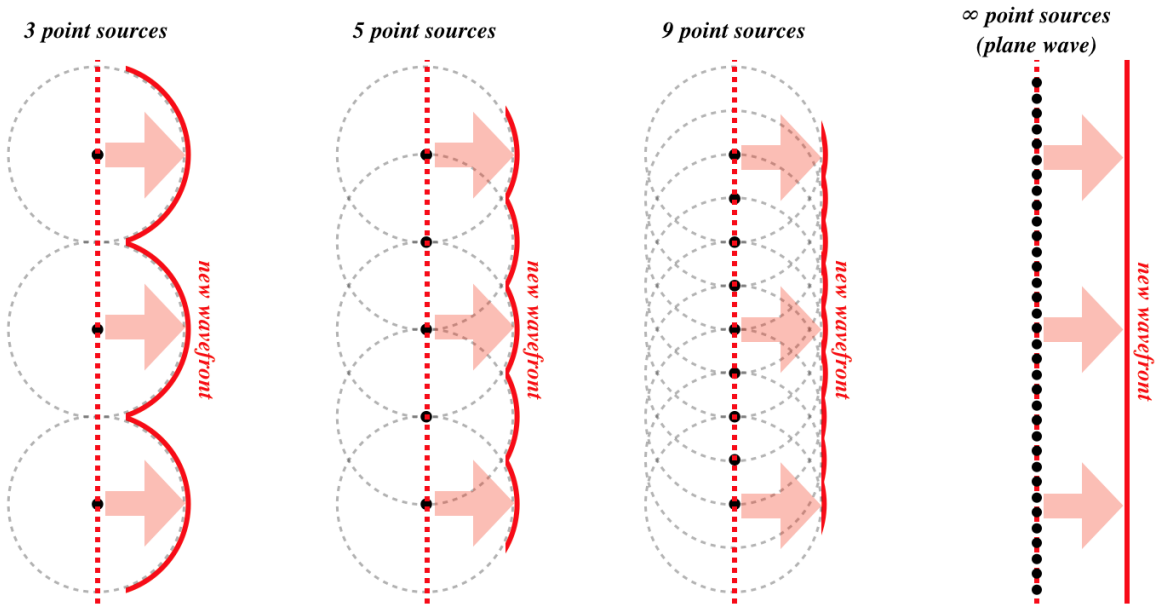
Huygens's Principle

When we discussed the case of a wave on a string, we said that the wave causes each particle on the string to vibrate up-and-down in harmonic motion. It should therefore not be surprising that if we grab the string at a single point and force it to vibrate in harmonic motion, that a wave will propagate away from that point. In fact, this gives us a way of describing how the wave propagates: The wave causes a single point to oscillate, which in turn causes a wave to be generated, which then vibrates another point, and so on. In the 17th century a Dutch scientist named Christian Huygens generalized this idea to three dimensions. The principle which now bears his name can be stated this way:

Every (3-dimensional) wave propagates by having every point on a wavefront being an independent generator of a new spherical wave, and the interference of all of those individual spherical waves results in the overall wave observed.

When we look at a single point light source, the farther away it is, the flatter the light wavefronts will be when they reach us. When the source is very far away (e.g. the sun), then the wavefronts are essentially flat. We call waves with such flat wavefronts *plane waves*, for obvious reasons. But now the question arises, “If Huygens’s principle is valid, how can plane waves occur?” After all, each point on the plane wave behaves as a point source of a spherical wave. Let's look at the spherical wave contributions of many point sources on a plane. We'll do this gradually, starting with just a few points on a plane, and filling in the spaces between them little-by-little:

Figure 2.1.2 – Plane Wave from Huygens's Principle

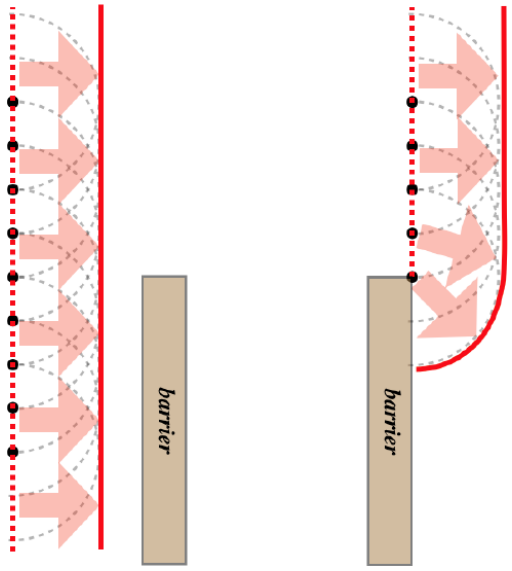


One might ask why a plane wave only propagates in a single direction. Suppose a plane wave propagating to the right. If each new wavefront becomes a source for a new wave, why don't waves come out of it in both directions? It is difficult to express in a simple diagram like the one above the effects of superposition, but the short answer is that there is destructive interference between all of the previous wavefronts and the new one, which results in zero wave energy traveling "backwards."

It should also be noted that a plane wave is a one-dimensional wave, which means that its intensity does not drop off with distance. But the intensities of the spherical wavelets do follow an inverse-square law. So if they get weaker with distance, why don't plane waves? The reason is that the farther a wavelet travels, the more other wavelets it encounters. These encounters result in constructive interference, bolstering the amplitude (and therefore the intensity) The rate at which the wavelets encounter other wavelets and constructively interfere is exactly enough to compensate for each wavelet losing its own individual intensity, maintaining the plane wave's intensity.

Where Huygens's principle becomes particularly useful is in explaining what happens when a plane wave encounters a barrier. A plane wave moves straight ahead because there is destructive interference of the wavelets in other directions. But a barrier removes a number of wavelets by either absorbing or reflecting the part of the wavefront from which those wavelets were going to spawn. The result is that the wave "bends around corners," a phenomenon known as *diffraction*.

Figure 2.1.3 – Diffraction from Huygens's Principle



Like other wave phenomena, this is not unique to light. Ocean waves diffract around barriers like reefs, peninsulas, and docks. It's certainly possible to hear a sound made from around a corner. It should be noted that the part of a wave that diffracts around a corner is no longer a plane wave, and is subject to the reduction in intensity the farther it travels. Of course reflections of waves are also responsible for their ability to change direction in the presence of barriers, but the phenomenon of diffraction in conjunction with interference leads to other important observable properties that we will deal with next.

Alert

You should be aware that diffraction is so intimately tied up with the interference effects that it causes (the subjects of the next few sections) that many physicists use the word "diffraction" to indicate the interference phenomena themselves, rather than the "going around corners" definition.

2.2: Double-Slit Interference

Splitting a Light Wave into Two Waves that Interfere

We now return to the topic of static interference patterns created from two sources, this time for light. As with sound, we first need to start with two light sources that are at the same frequency. In the case of light, we say that the sources are *monochromatic*. For sound we were able to keep track of the starting phases of sounds coming from separate speakers by connecting them to a common source, but for light it's a bit trickier. There simply isn't a way to coordinate the phases of light waves coming from two independent sources (like two light bulbs). Light waves from multiple independent sources have phases that are essentially distributed randomly, resulting in a variety of light referred to as *incoherent*. In fact, even light from a single source such as an incandescent bulb is incoherent, because the vibrations of the various electrons that create the waves are not coordinated. It turns out (for complicated reasons we won't go into) that after light travels a long distance the coherence of the waves grows (so light from the sun is highly coherent), but for experiments with light sources located here on Earth we are forced to use lasers, which do produce coherent light. Again, the reason that laser light is coherent is complicated, and outside the scope of this class.

Even with the coherence available from a single laser, we cannot coordinate the phases of two separate laser sources, so we need to somehow use the waves coming from a single laser source. We do this by directing the light from a single source through two very narrow adjacent slits, called a *double-slit apparatus*. Huygens's principle assures us that then each slit becomes a source for a spherical wave emanating from the position of each slit, and since the wavefront reaches each slit at the same time, the two sources start in phase, just like the tones coming from two speakers attached to the same source.

Okay, so to get an idea of the interference pattern created by such a device, we can map the points of constructive and destructive interference. We can do this by mapping what happens to two spherical waves that start at different positions near each other, and specifically keeping track of the crests (solid circles) and troughs (dashed circles). [*Note: The two waves shown are in different colors to make it easier to distinguish them – the actual light from both sources is all the same frequency/wavelength/color.*]

Figure 2.2.1 – Double-Slit Interference

A coherent plane wave comes into the double slit, and thanks to Huygens's principle, the slits filter-out only the point sources on the plane wave that can pass through them, turning the plane wave into two separate radial waves, which then interfere with each other. Whenever a crest meets a trough there is total destructive interference, and whenever two crests or two troughs meet, the interference is (maximally) constructive. We notice a number of things here:

- If we watch the points of total destructive and maximally constructive interference as the waves evolve, they follow approximately straight lines, all passing through the center point between the two slits.

- Because of symmetry, we see that these lines are symmetric about the horizontal line that divides the two slits, and that the center line itself is a line followed by a point of maximal constructive interference.
- These lines alternate in type as the angle increases – the central line is constructive, the lines on each side with the next-greatest angle trace points of destructive interference, the next pair of lines trace points of constructive interference, and so on.
- There are a limited number of these lines possible.

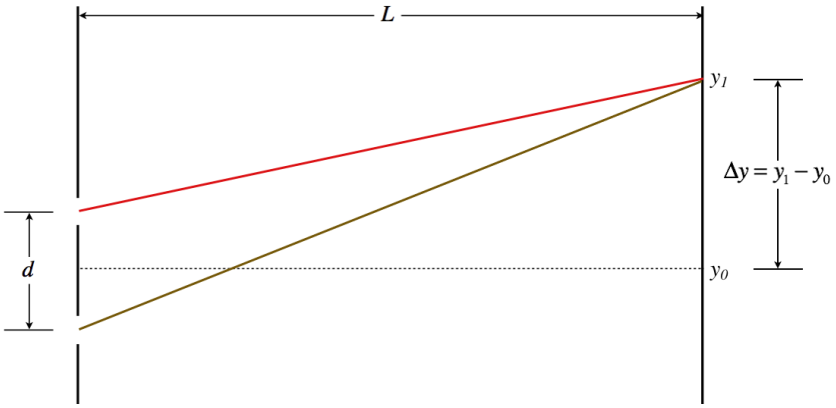
How are these effects *perceived*? Total destructive interference means zero intensity, which is the absence of any wave – darkness. Constructive interference is perceived as bright light, so if we placed a reflecting screen in the way of these light waves, we would see alternating regions of brightness and darkness, called *fringes*. It should be noted that the brightness varies continuously as one observes different positions on the screen, but we are focusing our attention on the brightest and darkest positions only. For the figure above, the screen would exhibit a *central bright fringe* directly across from the center point between the slits, then the first *dark fringes* some distance off-center, then more bright fringes outside of those. It is possible for a double-slit apparatus to produce either more or fewer fringes, depending upon the slit separation and the wavelength of the light. We will discuss the roles these variables play next.

Geometry of a Double-Slit Apparatus

Since we are (for now) only considering the brightest and darkest points, we can work with lines and geometry to get some mathematical answers. As stated above, these points only *approximately* follow straight lines from the center point, so our analysis will necessarily require some approximations. Whenever this is the case in physics, it is important to make a note of the physical features that go into determining the usefulness of the approximation as well as the tolerances we are willing to accept.

We begin by defining the slit separation (d) and the distance from the slits to a screen where the brightness interference pattern is seen (L). We also label some of the quantities related to the position on the screen in question.

Figure 2.2.2a – Double-Slit Geometry

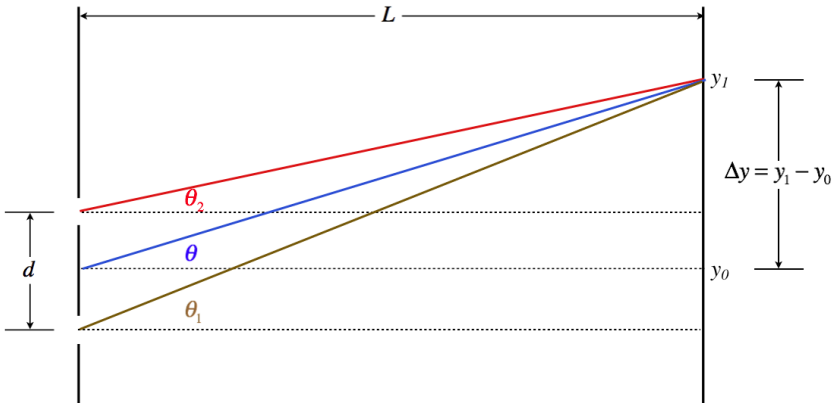


We are looking for those lines that define the destructive and constructive interference, so we want to express things in terms of a line that joins the midpoint of the two slits and the point located at y_1. In particular, we are looking for the angle θ that this line makes with the center line. We already know the center line traces a constructive interference, so our final answer should reflect this for $\theta = 0$.

The key physical argument we make here is that the wave that travels to y_1 from the upper slit has a shorter trip than the wave that gets there from the lower slit. The two waves start at the same time, and in phase, so this difference in distance traveled (Δx) accounts for the phase difference in the two waves that causes interference. So to relate the interference witnessed at y_1 to θ, we need to determine how (Δx) is related to θ.

As a start, we will draw in the line that goes from the midpoint of the slits to y_1, and label a bunch of angles:

Figure 2.2.2b – Double-Slit Geometry



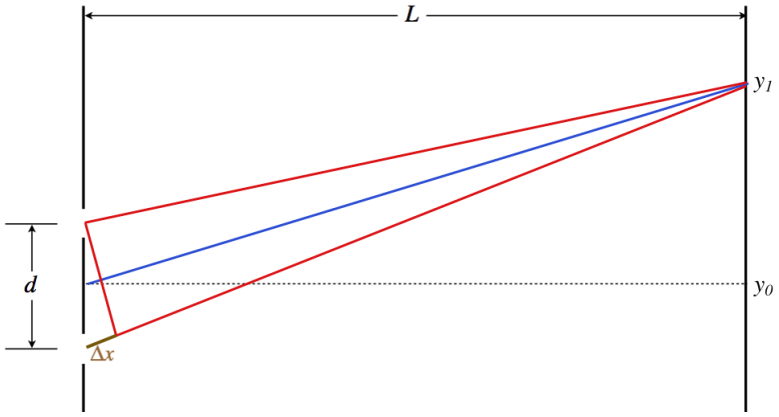
Now we need to do some math and apply some approximations. The tangents of these angles can be written in terms of the sides of the triangles they form:

$$\begin{aligned} \tan\theta_2 &= \frac{\Delta y - \frac{d}{2}}{L} \\ \tan\theta &= \frac{\Delta y}{L} \\ \tan\theta_1 &= \frac{\Delta y + \frac{d}{2}}{L} \end{aligned} \tag{2.2.1}$$

We don't actually require this math to convince us that if the slit separation is very small compared to the distance to the screen (i.e. $d \ll L$), then these three angles are all approximately equal. This is a good approximation, as this phenomenon is typically observed with slits separated by distances measured in fractions of millimeters, while distances to the screen are measured in meters. So henceforth we will make no mention of the angles θ_1 and θ_2.

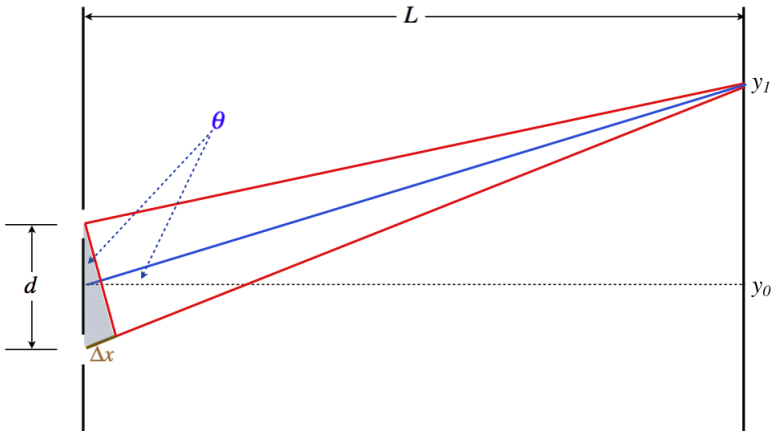
The next step is to break the lower (brown) line into two segments – one with the same length as the top (red) line that touches y_1 but doesn't quite reach the lower slit, and the other with the additional distance traveled, (Δx) that connects the first line to the lower slit. Then with the two equal-length segments, form an isosceles triangle:

Figure 2.2.2c – Double-Slit Geometry



Returning to our angle approximation where the top and bottom lines are approximately parallel, we see that this triangle has approximately two right angles at its base, which means there is a small right triangle formed by the base of the triangle, Δx, and the slit separation d. Imagine rotating the triangle clockwise. The angle at the top of this small triangle closes to zero at exactly the same moment that the blue line coincides with the center line, so this angle equals θ:

Figure 2.2.2d – Double-Slit Geometry



This gives us precisely the relationship between Δx and θ that we were looking for:

$$\Delta x = d \sin\theta \tag{2.2.2}$$

Now all we have to do is put this into the expression for total destructive and maximally-constructive interference. We know that total destructive interference occurs when the difference in distances traveled by the waves is an odd number of half-wavelengths, and constructive interference occurs when the the difference is an integer number of full wavelengths, so:

$$\begin{aligned} &\text{center of bright fringes:} && d\sin\theta = m\lambda \\ &\text{totally dark points:} && d\sin\theta = \left(m+\tfrac{1}{2}\right)\lambda \end{aligned} \qquad m = 0,\ \pm1,\ \pm2, \ldots \tag{2.2.3}$$

Let's take a moment to examine these equations, comparing what they require with the bulleted observations we made above:

- The plus-or-minus values of the integer m confirms that the fringes are symmetrically reflected across the center line.
- The case of $m = 0$ for constructive interference corresponds to the center line.
- Moving out from the center, the next fringe of any kind occurs when $m = 0$ for destructive interference. Then the next occurs for $m = 1$ for constructive interference, and so on – the bright and dark fringes alternate.
- Not all integer values of m will work, because the absolute value of $\sin\theta$ can never exceed 1. When the absolute value of m gets too high, this relation cannot possibly hold, placing a limit on the number of fringes. This limit is determined by the ratio of the wavelength to the slit separation.

It is sometimes useful to convert this result into measurements of distances from the center line on the screen, rather than the angle θ. To get this, we need the distance L, which was not necessary for the solution above (other than assuming it is much larger than d). Calling the distance from the center line to the m^{th} fringe y_m, we use the fact that the tangent of the angle is the rise over the run ($y_m = L\tan\theta_m$) to get:

$$\begin{aligned} &\text{center of bright fringes:} && y_m = L\tan\left[\sin^{-1} m\frac{\lambda}{d}\right] \\ &\text{totally dark points:} && y_m = L\tan\left[\sin^{-1}\left(m+\tfrac{1}{2}\right)\frac{\lambda}{d}\right] \end{aligned} \qquad m = 0,\ \pm1,\ \pm2, \ldots \tag{2.2.4}$$

So long as we are careful, we can simplify this with a second approximation. If the *angle is small*, then the tangent and sine of that angle are approximately equal. This simplifies the above result to:

$$\text{for small } \theta: \qquad \begin{aligned} &\text{center of bright fringes:} && y_m = m\frac{\lambda L}{d} \\ &\text{totally dark points:} && y_m = \left(m+\tfrac{1}{2}\right)\frac{\lambda L}{d} \end{aligned} \qquad m = 0,\ \pm1,\ \pm2, \ldots \tag{2.2.5}$$

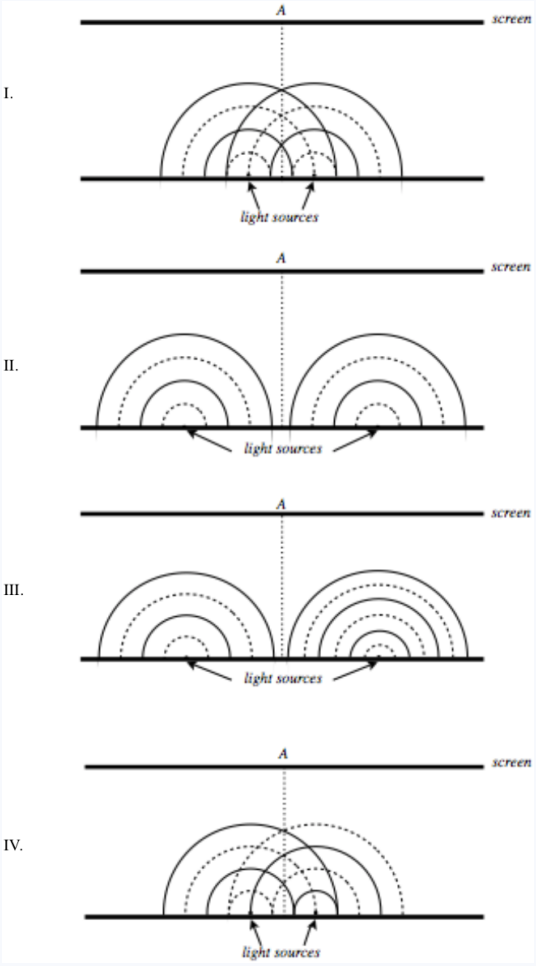
This shows us that for small angles, fringes of the same type are equally-spaced on the screen, with a spacing of:

$$\Delta y = \frac{\lambda L}{d} \tag{2.2.6}$$

Example 2.2.1

Below are four depictions of two point sources of light (not necessarily caused by two slits), using the wave front model. These depictions are "snap shots," 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?*



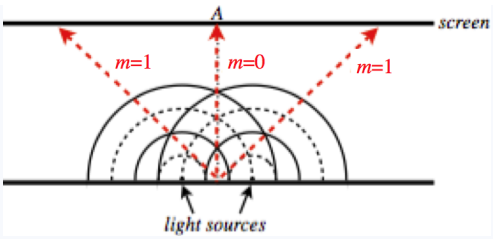
Solution

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

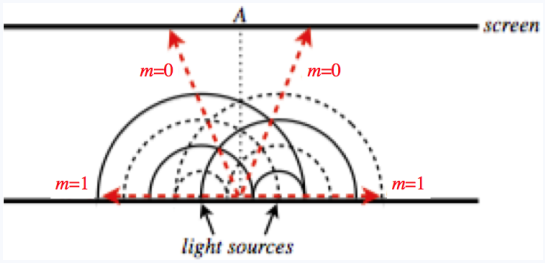
c. N/A

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.

Double Slit Intensity Pattern

The answers above only apply to the specific positions where there is totally destructive or maximally constructive interference. What about the points in between? For this answer, we return to Equation 1.4.10, which relates any phase difference of two waves to the intensity of the wave in comparison to its maximum intensity (when maximal constructive interference occurs). As noted earlier, the only source of phase difference is the distance traveled by the two waves, so:

$$\left.\begin{aligned} I &= I_o\cos^2\left(\frac{\Delta\Phi}{2}\right) \\ \Delta\Phi &= \frac{2\pi}{\lambda}\Delta x \\ \Delta x &= d\sin\theta \end{aligned}\right\} \quad\Rightarrow\quad I(\theta) = I_o\cos^2\left[\frac{\pi d\sin\theta}{\lambda}\right] \tag{2.2.7}$$

It's easy to see that this works correctly for the specific cases of total destructive and maximal constructive interference, as the intensity vanishes for the destructive angles, and equals I_o for the constructive angles. If the angle is small, then we can approximate this answer in terms of the distance from the center line:

$$I(y) = I_o \cos^2\left[\frac{\pi y d}{\lambda L}\right] \tag{2.2.8}$$

Activity

To see all the features of double-slit interference, check out this simulator. To simulate double slit interference for light, take the following steps:

1. Select and click on the "Interference" box.
2. In the control box, click the laser icon:  .
3. In the control box, click the "Screen" toggle box to see the fringes.
4. Click on the green buttons on the lasers to start propagating the light waves.
5. In the control box, you can adjust frequency and slit separation to see the effects on the interference pattern.
6. You can click on the intensity toggle box in the control box to see the graph of the intensity at the screen, as described by Equation 2.2.8.
7. You can even explore this phenomenon *quantitatively* (i.e. check the math derived above) by reading the slit separation in the control box, dragging measuring devices from above the control box, and pausing the simulation.

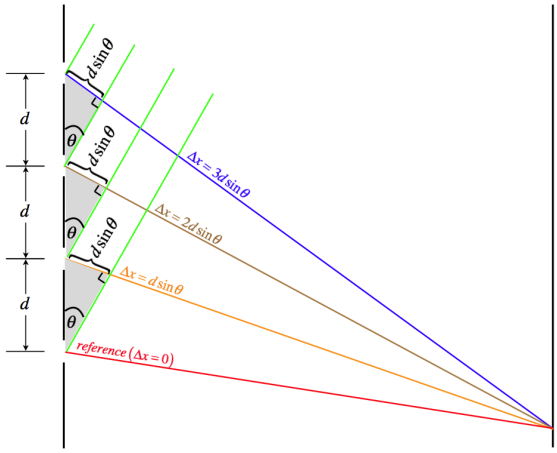
2.3: Diffraction Gratings

Adding More Slits

After having determined the interference pattern associated with two slits, it makes one wonder what would happen if many more (equally-spaced) slits are added. We can recycle our geometrical analysis from the double slit problem to answer this question. Let's look at the example of four slits.

We begin once again with the assumption that the distance to the screen is significantly larger than the separation of adjacent slits: $d \ll L$). Starting with the lowest slit of the four as a "reference" and repeating the double-slit geometry for each slit going up from there, we have a diagram that looks like this:

Figure 2.3.1 - Geometry of Four Slits



The Δx in each case is the difference in distance traveled compared to the reference slit. So the extra distance traveled by the wave following the blue path is three times as great as the extra distance traveled by the wave following the orange path.

Alert

This diagram is blown-up for clarity, but doing so makes the angles quite different from each other. With the proper scale in place the approximations of equal angles (and equal Δx's throughout) would be more apparent.

Okay, so as our first task, we will look for the position where the first bright fringe is located. For this to occur, we need all four waves to be in phase, which means that Δx has to be a full wavelength, giving us the same formula for bright fringes that we found for the double slit:

$$d \sin\theta = m\lambda \ , \qquad m = 0, \ \pm 1, \ \pm 2, \ \ldots \tag{2.3.1}$$

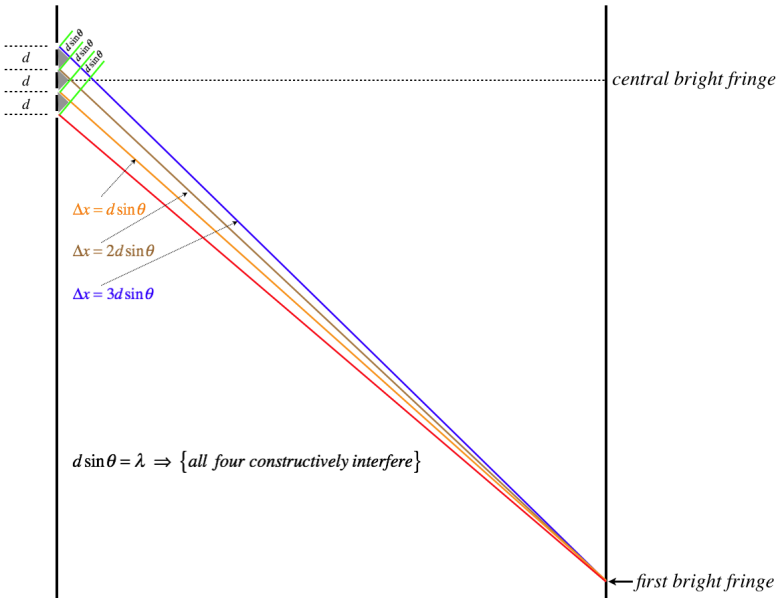
[*It should be noted that the positions of the fringes on the screen are measured from the horizontal line passing through the center of the collection of slits, as we did with the double slit.*]

Does this mean that the result for several slits is identical to that of the double slit? Certainly not! First of all, there are many more sources of light, all interfering constructively, which means that the bright fringes are much brighter. How much brighter? Well, with four slits, as in the example here, the amplitude of a single slit is multiplied by 4, making the intensity (which goes as the square of the amplitude) 16 times greater than a single slit. For the double slit, the intensity was increased by a factor of 4 (the amplitude was doubled). Therefore doubling the number of slits increased the intensity of the bright fringes by a factor of 4. But wait, doubling the number of slits only lets in twice as much energy per second, so how is the intensity increasing so much?

The answer to this puzzle involves how *concentrated* the bright fringes are. All bright fringes have a point of maximum brightness that tapers down to the dark fringes. If the rate at which the brightness tapers down is greater, then the brightness (energy density) near those maximum points can go up, and the energy density near the dark fringes goes down, such that the same total energy hits the screen. But it turns out there is even a little more to it than this, as we will now see.

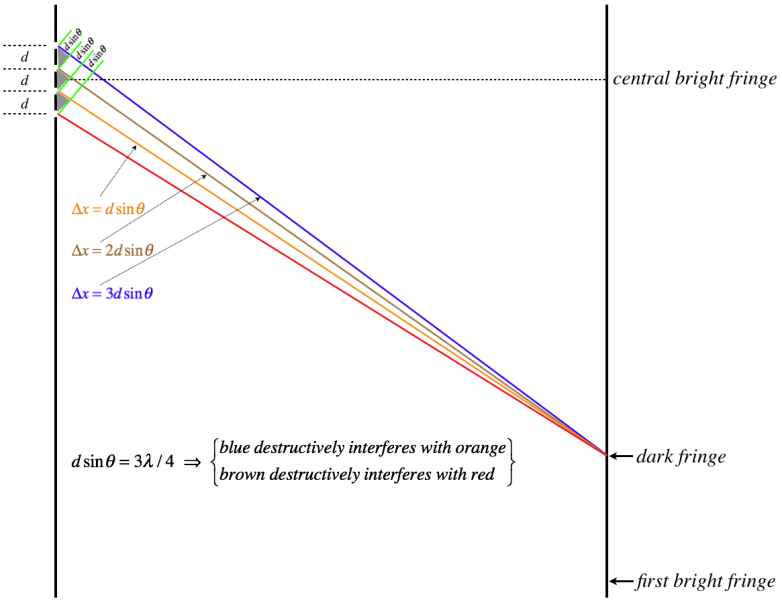
To demonstrate this phenomenon, it becomes necessary to redraw the figure above a little closer to the actual scale. We of course cannot possibly get very close to the actual scale, as slit separations are typically fractions of millimeters, while distances to screens are usually tens or hundreds of centimeters, but we will use what space we can manage. As before, we will use the red line as the reference, and compare the distances traveled by the other three light waves.

Figure 2.3.2a - Finding Dark Fringes



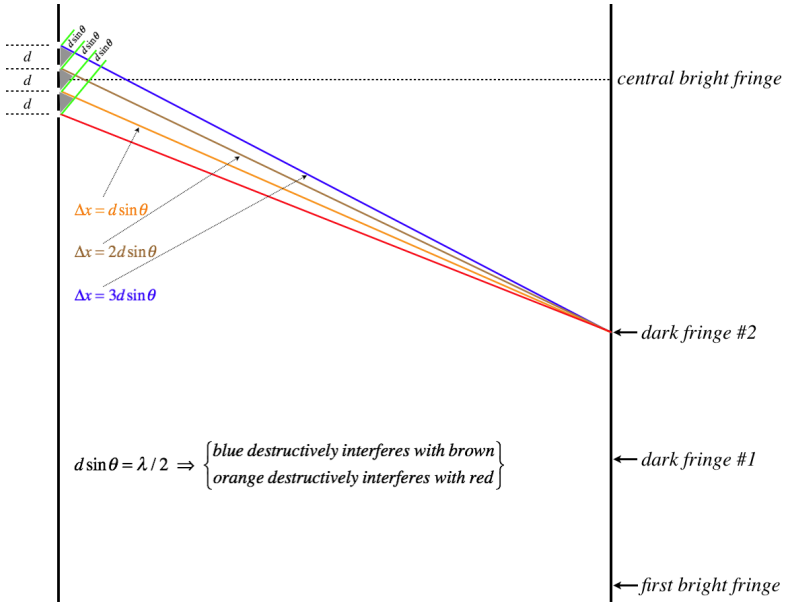
We'll start with the bright fringe, and start working our way closer to the central bright fringe until we hit a dark fringe. Strangely, we find that the first position of total destructive interference we encounter does *not* occur at the halfway point, as it did for the double slit! Note that when the distance $\Delta x = d\sin\theta$ equals *three-quarters of a wavelength*, then the wave that follows the blue path will travel 1.5 wavelengths farther than the wave that follows the the orange path, and as this is an odd number of half wavelengths, these waves will cancel. The same is true for the waves that follow the brown and red paths, which means that position will be completely dark.

Figure 2.3.2b - Finding Dark Fringes



So what happens if we keep going up the screen? We don't find any more maximally-bright fringes (all four waves can't be in phase), but we *do* find another totally dark position. It occurs when the distance $\Delta x = d\sin\theta$ equals *one-half of a wavelength*. In this case, The wave that follows the blue path travels one half-wavelength farther than the wave that follows the brown path, and the waves that follow the orange and red paths also differ in the distance they travel by one half wavelength. So the blue path and red path waves cancel, as do the brown path and yellow path waves, resulting in total darkness.

Figure 2.3.2c - Finding Dark Fringes



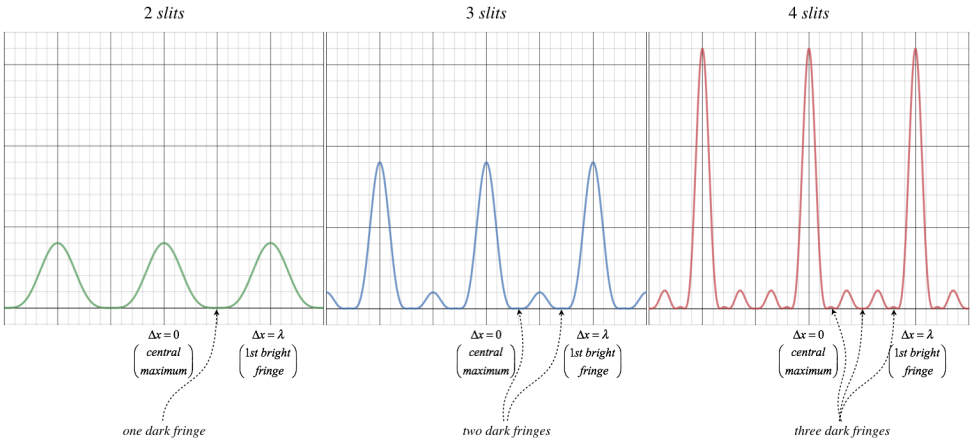
There is one other time when a dark fringe occurs. This happens when the distance $\Delta x = d\sin\theta$ equals *one-quarter of a wavelength*. Once again, alternate slits interfere with each other, as the waves travel distances that differ by a half-wavelength.

We can also show this phenomenon mathematically, by superposing (adding) the wave functions. The waves start in phase at the slits, so all of the phase constants are equal (and we choose them to be zero at $t = 0$), so all that remains of the wave functions is the position dependence. Once again, all that matters are the *differences* in distances traveled with the reference slit (whose difference with itself is zero), so the superposition intensity looks like:

$$\begin{aligned}
I_{2\,\text{slits}}(\theta) &= \{f_1(\theta)+f_2(\theta)\}^2 &&= \left\{\cos 0+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)d\sin\theta\right]\right\}^2 \\
I_{3\,\text{slits}}(\theta) &= \{f_1(\theta)+f_2(\theta)+f_3(\theta)\}^2 &&= \left\{\cos 0+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)d\sin\theta\right]+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)2d\sin\theta\right]\right\}^2 \\
I_{4\,\text{slits}}(\theta) &= \{f_1(\theta)+f_2(\theta)+f_3(\theta)+f_4(\theta)\}^2 &&= \left\{\cos 0+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)d\sin\theta\right]+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)2d\sin\theta\right]+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)3d\sin\theta\right]\right\}^2 \\
I_{n\,\text{slits}}(\theta) &= \{f_1(\theta)+f_2(\theta)+\cdots+f_n(\theta)\}^2 &&= \left\{\cos 0+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)d\sin\theta\right]+\cdots+\cos\left[\left(\tfrac{2\pi}{\lambda}\right)(n-1)\,d\sin\theta\right]\right\}^2
\end{aligned} \tag{2.3.2}$$

Putting these functions into a graphing calculator confirms what we found above, as well as what we suspect about n slits – that there are $n-1$ dark fringes between each maximally-bright fringe.

Figure 2.3.3 - Comparison of Interference Patterns by Number of Slits



Notice that the bright fringes for any number of slits occur at the same places as for the double slit (provided they have the same slit separation), and that the number of dark fringes between bright fringes goes up by one every time another slit is added. Also notice that the maximum intensity of the double slit is 4 units, the 3-slit case has a maximum intensity of 9 units, and for 4-slits it is 16 units, as we expect when the amplitude increases by one unit with the addition of each slit. But also notice that the *widths* of the bright fringes get narrower, indicating that the energy becomes more concentrated near the brightness maxima, and less concentrated near the dark fringes.

It turns out that we can mathematically check that the energy is in fact conserved by this mechanism. Recall that the intensity is related to power *density*, which means that if we integrate one of these curves over a full interval of space that the light is landing (say, between adjacent bright maxima), we get a measure of the energy landing in that region per unit time. Once again the graphing calculator comes in handy (unless integrating the intensity functions above is your idea of fun) as areas under these curves between maxima come out to be in relative proportions of 2:3:4 – the total energy landing on the screen every second really is proportional to the number of slits allowing light through!

Adding Many, Many More Slits

We know that the regions where the bright fringes peak get more concentrated light, and that there are more dark fringes between them when the number of slits is increased. One can imagine that in the limit where very many slits are used (a device called a *diffraction grating*), the result is very sharp, very bright lines lines at the points of maximum constructive interference, and darkness everywhere else. As we will see, this will be an extremely useful feature. But there is one assumption we have made here that needs to be emphasized. Because d is so small compared to the distance to the screen, it was easy to ignore the fact that this particular calculation required the assumption that the first bright fringe be farther from the center line than the outermost slit (we assumed that the wavelength was long enough that this had to be true). So creating a sharper interference pattern for a given wavelength of light by adding more slits at the same separation on both sides of the center line has limitations, because when the number of slits gets very large, the added slits go past the bright fringe. However, if more slits are added by *squeezing them closer together* (making d smaller), then for a given wavelength, then not only are there more slits, but the angle to the first bright fringe increases, thanks to the relation $d\sin\theta = m\lambda$.

It is for this reason that diffraction gratings are generally characterized by their *grating density* – the number of slits per unit distance. Of course such a number can be converted into a slit separation: If a diffraction grating has a grating density of 100 slits per cm, then the slits must be separated by $d = \frac{1}{100}cm = 10^{-4}m$. This number can then be used in calculations for the angle at which bright fringes are seen.

It should also be mentioned that like double slits, diffraction gratings do allow for more than one bright fringe (as before, depending upon the ratio of d and λ). For a typical double slit experiment, the goal is usually to show a broad interference pattern – many fringes. If the slit separation is too small, then the angles between the fringes are large, resulting in very few fringes, widely separated, foiling the goal of such an experiment. But use of a diffraction grating has a different goal (very sharp bright fringes), which requires that the slits be separated by much smaller distances. This results in far fewer fringes, separated by large angles. So while the calculation for the angles of bright fringes is the same for both devices, for a given range of wavelengths, their slit separations are usually quite different.

Applications of Diffraction Gratings

It was stated above that sharp bright fringes are very useful in applications. To see why this is so, suppose one wishes to use a diffraction device to measure the wavelength of a monochromatic light. This is straightforward – shine the light through any number of slits with a known slit spacing, and measure the angle at which the first bright fringe is deflected from the central bright fringe, then plug into $d\sin\theta = m\lambda$ (with $m = 1$) and solve for λ. The only real challenge to this procedure is measuring the angle. Of course, if we shine the light onto a screen whose distance we know from the slits, we can measure the distances between the bright fringes, and compute the angle from there. But *still* we have a problem if we want to be precise. If a double-slit is used, then the bright fringe is rather broad, and it might be challenging to get a good measurement of its center. With a diffraction grating, the bright fringe is much better defined. Furthermore, the light we are looking at may not be very intense, and a diffraction crating lets much more of the light in, and the bright fringe is much easier to see than it would be for a double-slit.

But even these two advantages pale in comparison to the third. We have not yet considered what happens if we look at light that is not monochromatic. Suppose the incoming light is a mix of three or four colors. The separate colors don't interfere in a static manner with each other (they can create "beats," but the frequency differences for light are so great that these will not be observable) they only observably interfere with themselves. As such, a beam of light with three colors will exhibit three separate interference patterns when passed though a single device (i.e. they all experience the same slit separation). The wave with color corresponding to the shortest wavelength will have its first bright fringe deflected by the smallest angle. If this light is passed through a double-slit, the interference patterns blend with each other, making it hard to separate the component colors. But a diffraction grating makes three sharp, distinct, first-order bright fringes, making it easy to determine the constituent colors of the incoming light.

An important part of the fields of chemistry and astronomy is the method of measurement called *spectroscopy*. In Physics 9D, you will learn that matter emits and absorbs light in very peculiar ways. You might think that electrons in atoms can vibrate at any frequency at all and therefore emit or absorb a nice, smooth continuous spectrum of light, but it turns out that they cannot. In fact each atom has a unique “fingerprint” of specific frequencies of light that it emits and absorbs. This means that when light emitted from a certain substance is passed through a diffraction grating, this fingerprint is manifested as a specific set of bright fringes (called *spectral lines*). This means that we can ascertain from a distance (in the case of astronomy, very great distances!) the composition of the matter that is emitting light. These fingerprints are so specific and unique that even if several different substances are emitting light, they can generally be sorted out.

One might worry that since stars are moving relative to the earth, that we might get the elements wrong, since what we will see in the *spectrometer* (a device with a diffraction grating) will measure doppler-shifted wavelengths. But it isn't the exact positions of the spectral lines that tells us the elements emitting the line, but rather their *relative* positions. That is, every spectral line is doppler-shifted, so the "barcode" essentially looks the same for hydrogen regardless of its relative motion, because the whole barcode is just shifted toward longer wavelengths if it is moving away from the spectrometer, and toward the shorter wavelengths if moving toward the spectrometer.

But astronomers can do even more than identify elements in burning stars. We know what the barcode for hydrogen looks like when the source is at rest relative to the spectrometer, so when we see the hydrogen barcode pop up for a star, we can measure how much the barcode in the spectrometer is shifted compared to the stationary case, and we can use the amount of shift to determine how fast the star is moving relative to earth!

2.4: Single-Slit Diffraction

Slits Are Not Actually Point Sources

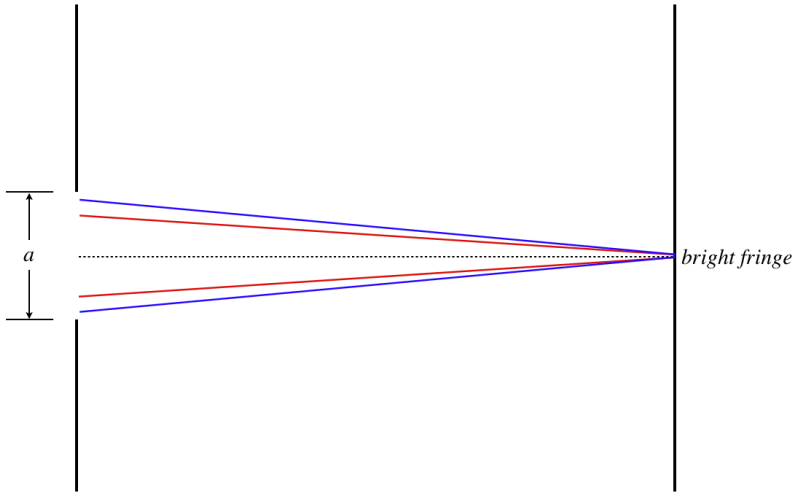
In our discussion of the double slit and diffraction grating, we made the assumption that the gaps that we call slits are so narrow that they can essentially be treated as point sources, making the analysis using Huygens's principle simple to do. But in reality we know that these gaps do not have infinitesimal width, and we need to consider what happens to the light when the approximation of "very thin gaps" breaks down. To do so, we will not consider a grating, or even a double-slit; we'll look at the effect that a single slit of a measurable gap size has on the light that passes through it. Notice that whatever this effect might be, when we extend the result to two or more slits, the effect will occur for every slit, superimposing itself on the multiple-slit interference pattern. But we are getting ahead of ourselves...

We already know that a plane wave passing through a single slit will diffract around the corners, so it will not simply leave a single bar of light on the screen the thickness of the gap – it will spread out. But what else can we say about it? Well, we know that without the aperture, all the Huygens wavelets would continue interfering perfectly to continue the plane wave, but when the portions of the plane wave outside the aperture are excluded, the effects of interference between wavelets is bound to change. We will analyze the effect by essentially following the procedure for many (infinite number) of thin slits that are infinitesimally close together.

Single Slit Interference Pattern

Let's call the gap width of the aperture a, and assume that this is much smaller than the distance to the screen, as in the figure below. We then consider what happens to the wavelets originating from every point within this region. When we look at how the screen opposite a single slit is illuminated, on the screen at the center line we observe a brightness maximum. You can think of such a situation as an infinite number of double-slits that are split by the center line with different slit separations. For every wavelet above the center line, there is a "twin" wavelet on the opposite side of the center line that travels the same distance to the screen (depicted by lines of the same color in the figure below), resulting in constructive interference. Of course, the fact that pairs constructively interfere with each other does not guarantee that the result of two constructively-interfering wavelets will not cancel with two other constructively-interfering wavelets (i.e. one pair creating a doubly-high peak, and the other a doubly-deep trough). In fact this can happen, but if it does, it's only for select wavelets – it can't persist for the entire aperture and leave darkness at the center line. Without going into the math, wavelets find it exceedingly difficult to find canceling partners at the center line, and on balance the interference is highly constructive – the center line is the brightest point in the entire interference pattern.

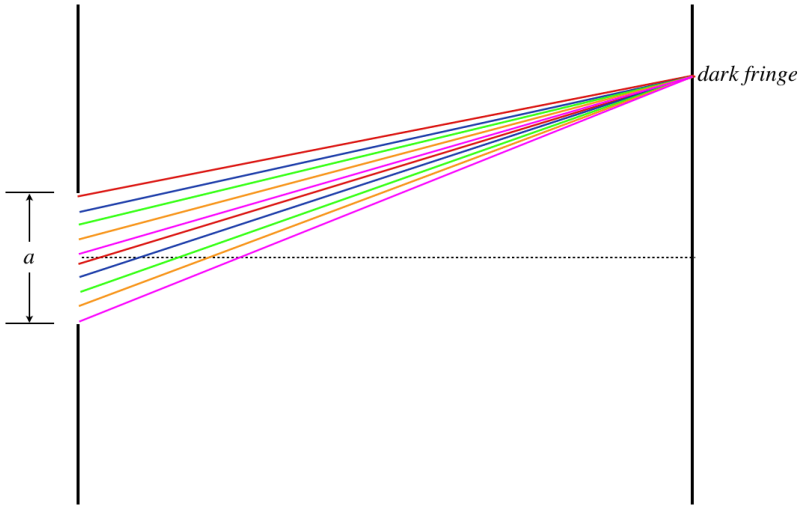
Figure 2.4.1 - Wavelet Pairs Constructively Interfere at the Center Line



Okay, so what about dark fringes – will we see these on the screen? Yes! To see why, we will once again find pairs of wavelets on both sides of the center line, which in this case travel different distances to the screen, differing by one-half wavelength for the first dark fringe. For this case, we pair-off the wavelet originating at the top of the slit with the wavelet originating just below the center line, and continue pairing them as we go down, until the wavelet at the bottom edge pairs with the wavelet originating just above the center line. This is depicted in the figure below with pairs of lines of the same color. The difference in distances for these pairs

will all be the same ($d\sin\theta$, where in this case d is actually $\frac{a}{2}$), and when this difference is one-half wavelength, they all cancel each other pairwise, leaving a dark fringe.

Figure 2.4.2 - Wavelet Pairs Destructively Interfering at the First Dark Fringe



Note that the same geometry holds below the center line as well. Setting the extra distance traveled by the twin wavelets equal to a have wavelength, we get the angle of the first dark fringe:

$$\text{first dark fringe:} \quad \frac{a}{2}\sin\theta = \frac{\lambda}{2} \quad \Rightarrow \quad \sin\theta = \pm\frac{\lambda}{a} \tag{2.4.1}$$

As we move upward on the screen, wavelets will again find their destructive twins and create dark additional dark fringes. It is a bit tricky for us to find the second dark fringe, however. The natural approach is to assume that the next dark fringe occurs when the pairs shown above travel distances that differ by three half-wavelengths, giving the result $\sin\theta = \pm 3\frac{\lambda}{a}$. But in fact this result incorrectly skips the second dark fringe, and goes to the third! To see why, we note that we can pair-off wavelets in a way other than across the center line. Specifically, we can think of this single slit as *two adjacent* single slits, one that has the center line as its lower edge, and one that has the center line as its upper edge. In this case, the wavelets pair-off within the top half, and then again within the bottom half separately. In this case, the only change in the math involves replacing $\frac{a}{2}$ with $\frac{a}{4}$, which means the second dark fringe satisfies:

$$\text{second dark fringe:} \quad \frac{a}{4}\sin\theta = \frac{\lambda}{2} \quad \Rightarrow \quad \sin\theta = \pm 2\frac{\lambda}{a} \tag{2.4.2}$$

We can similarly break the slit into three separate slits, which changes the separation of the starting wavelets to $\frac{a}{6}$, and increments the constant in the formula to 3. For the m^{th} dark fringe, we therefore have:

$$m^{th}\text{ dark fringe:} \quad \sin\theta = \pm m\frac{\lambda}{a} \tag{2.4.3}$$

The bright fringes only approximately follow the same spacing pattern, not exactly located halfway between the dark fringes, but using the pairwise approach doesn't tell us much about the intensity of those bright regions, for the same reason it didn't for the central bright fringe – constructive pairs will not be in phase with other constructive pairs. Significantly more math is required to deal with the intensity of the bright fringes.

Intensity

To compute the intensity of the interference pattern for a single slit, we treat every point in the slit as a source of an individual Huygens wavelet, and sum the contributions of all the waves coming out at an arbitrary angle. One way to think of this is to go back to the diffraction grating case, expressed in Equation 3.3.2. With the slit being completely open, however, the space between the slits (d) goes to zero, and the number of slits (n) goes to infinity. There is of course more to the calculation than this, and either the calculus or the "phasor method" described by many standard physics textbooks will reach the famous result below, and the reader is encouraged to have a look at these derivations. But these derivations do not contribute to the understanding of this

phenomenon, nor are they procedures essential to a wide range of future physics calculations, so we will omit them here, and jump to the end result.

If we define the amplitude of the total wave on the center line to be A_o due to the superposition of all the wavelets, then the amplitude of the wave at an angle θ off the center line is given by:

$$A(\theta) = \frac{\lambda A_o}{\pi a \sin\theta} \sin\left(\frac{\pi a \sin\theta}{\lambda}\right) \tag{2.4.4}$$

Yes, you are reading that right, there is a sine function of θ *within* another sine function. This is often written more succinctly by defining a new variable that is an implicit function of θ:

$$A(\alpha) = A_o \frac{\sin\alpha}{\alpha}, \qquad \alpha(\theta) \equiv \frac{\pi a}{\lambda}\sin\theta \tag{2.4.5}$$

This function comes up frequently enough in math and physics that it has even been given its own name – it is sometimes referred to as a *sinc function*.

Alert

It is important to understand that this expression compares the amplitude at various angles to the amplitude on the center line, equal (or approximately equal) distances from the slit. It does not provide a comparison of the amplitude of the light wave after passing through the slit to the amplitude of the plane wave before it enters the slit.

We know that the intensity of the wave at the center line is proportional to the square of the amplitude there, and that the intensity of the wave at an angle with the center line is proportional to the square of the amplitude there, and that the constants of proportionality are the same in both cases, so we immediately have a comparison of intensities:

$$I(\alpha) = I_o \left[\frac{\sin\alpha}{\alpha}\right]^2 \tag{2.4.6}$$

If the angle θ happens to be small, then α can be written as a function of distance y from the center line on the screen, as we did in Equation 3.2.5 for the double slit, giving:

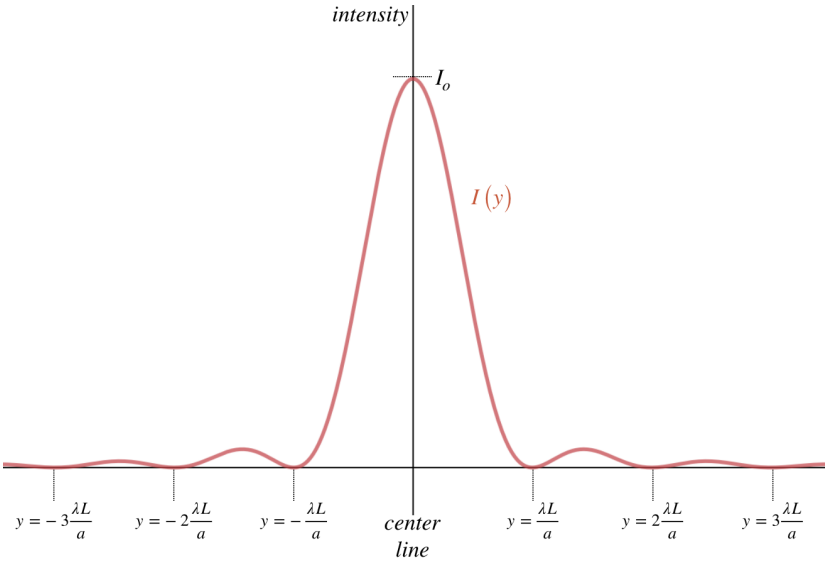
$$\alpha(y) \equiv \frac{\pi a y}{\lambda L}\ , \tag{2.4.7}$$

where, as before, L is the distance from the slit to the screen.

Perhaps you are concerned about the behavior of this function at the center line? After all, the value of the function α there does vanish, and this function appears in the denominator. But the numerator also vanishes at the center line, and L'Hôpital's Rule saves the day, giving the sinc function a value of 1 for $\alpha = 0$, resulting in the intensity equaling I_o, as it should.

A graph of the intensity of the full interference pattern looks like this:

Figure 2.4.3 - Single Slit Diffraction Intensity



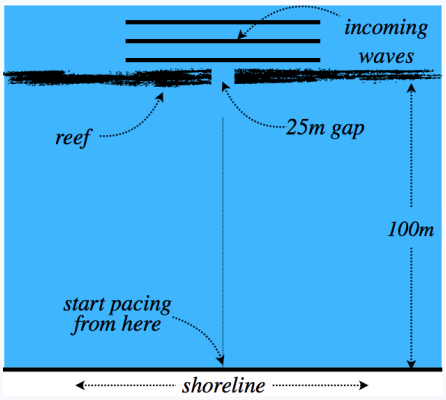
Let's point out a few of the more prominent features of this intensity pattern.

- The dark fringes are regularly spaced, in exactly the manner described by Equation 2.4.3 (note: $\sin\theta \approx \frac{y}{L}$).
- The central bright fringe has an intensity significantly greater than the other bright fringes, more that 20 times greater than the first order peak.
- Using calculus to find the placement of the non-central maxima reveals that they are not quite evenly-spaced – they do not fall halfway between the dark fringes.

We have assumed for simplicity the geometry of a long rectangular slit. If we were instead shining the light through a circular hole, this pattern would occur in every direction of two dimensions, resulting in concentric bright and dark circles, rather than fringes.

Example 2.4.1

You are on a sunny Hawaiian beach, trying to relax after a grueling quarter of Physics 9B. You would like to recline in your beach chair with your feet in the water, but don't want to get crushed by shore break while you snooze. About 100 meters off shore, you see an exposed reef that acts as a breakwater, but there is a gap in it, and waves (whose crests are parallel to the shore) are coming through that gap. While watching the waves, you see a surfer paddle out through the gap, and you use the perspective this event affords you to estimate that the gap is 25 meters wide (the diagram below is not to scale). You time a wave as it comes from the reef, estimating that it takes about 2 minutes for a wave to get to the shore from the gap, and the waves hit the shore roughly every 7 seconds. Starting from the point on the beach directly in line with the center of the gap, roughly how many paces (each pace being 1 meter in length) must you walk along the beach so that you can plant your beach chair and get the minimum wave intensity?



Solution

This is a problem in single-slit diffraction, where we are searching for the first "dark fringe" (place where destructive interference occurs). We can use Equation 3.4.3 for finding the angular deviation from the center line for a single slit, but it requires the wavelength of the wave as well as the slit gap. We have the latter, but we need to calculate the former. We can determine the wave speed and we are given the period, so:

$$\lambda = vT = \left(\frac{100m}{120s}\right)(7s) = 5.83m$$

Now we can plug this wavelength into Equation 3.4.3 to find the angle of the first dark fringe:

$$\sin\theta = \frac{\lambda}{a} \quad \Rightarrow \quad \theta = \sin^{-1}\left(\frac{5.83m}{25m}\right) = 13.5^o$$

The distance from the gap to the shoreline and the angle are known, so we can determine how far along the shore the dark fringe hits:

$$y = x\tan\theta = (100m)\tan 13.5^o = 24m$$

So you need to walk 24 paces.

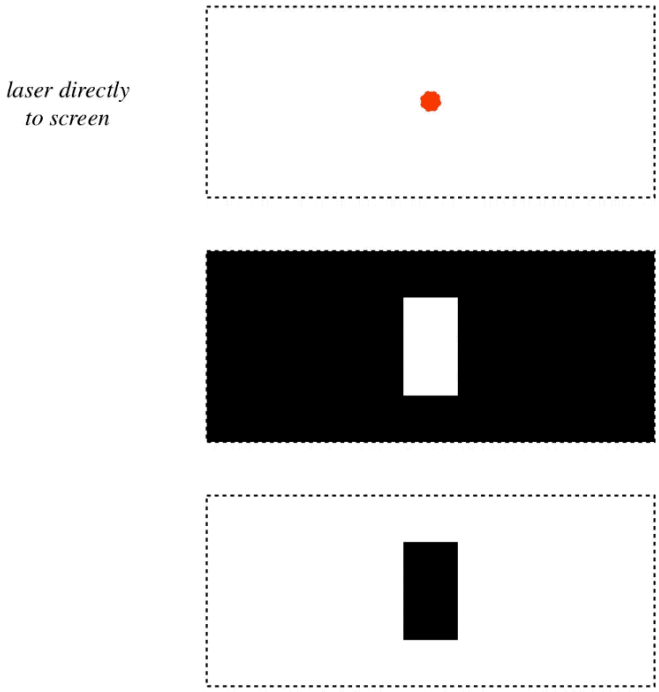
You might be tempted to use the "small angle" equation to solve this more directly, and in fact the angle is quite small. But we have defined our measurement limits in terms of paces, and using the small angle formula we end up with an answer of 23 paces, so while the approximation is very good (it is only off by less than 5%), even our rather coarse measurement scheme notices the difference. [Okay, so "notice" might be too strong a word, as the wave intensity one pace from a minimum and 23 paces from the maximum is not going to be significant.]

Diffraction for an "Inverse Slit"

We can use our knowledge of waves to determine the light pattern we will see when the incoming plane wave diffracts around a thin barrier. Imagine starting with a plane barrier, out of which we cut a tiny sliver. Described above is what we see if coherent light is shone through the opening we have created in the barrier, but what if we shine the same light on *just the sliver*? That is, instead of only allowing light to pass through a thin space, we let the light pass everywhere *except* the thin space.

Imagine a tight laser beam in three different situations: First, it goes straight to the screen unimpeded. As with all laser beams, it spreads very little during its journey. Second, it encounters a thin slit that is a little bit smaller than the width of the beam. Naturally a single-slit diffraction pattern appears on the screen. And third, the beam encounters only a sliver that has the same dimensions as the single slit, so that the outer edges of the beam go past the edges of the sliver. Our question is what happens in this third case.

Figure 2.4.5 - Three Laser Beam Results



If the light from the second case was allowed to superpose with the light from the third case, it should be pretty clear that the result will be the first case. But for the second case, some light lands outside the beam's confines (thanks to diffraction), which means that for the superposition to occur, the third case must also send light to those outer regions with exactly the same amplitudes as the slit, though the light from the sliver must be π radians out of phase with the light from the slit. But if the sliver is by itself, the light it sends outside the beam region doesn't cancel with anything, which means it shows up on the screen. The end result is that the interference pattern outside the beam region must be the same for the sliver as it was for the slit.

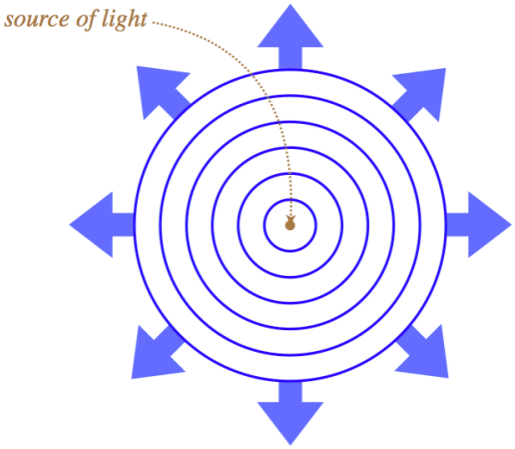
What about the central bright fringe? For a single slit, the central maximum is not as bright as the unimpeded beam (because some of the light energy is diverted by diffraction). For the superposition to apply, this means that the region directly behind the sliver must also be illuminated. The relative brightness of the central maximum with the outer fringes may be different for the slit and the sliver, but the fringe spacings are the same in both cases, giving essentially the same diffraction patterns for both cases. This phenomenon is known as *Babinet's principle*.

2.5: Reflection and Refraction

Rays

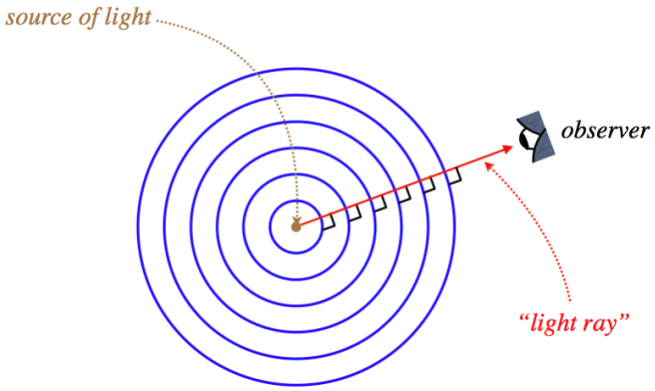
As we consider more phenomena associated with light, one of our primary concerns will be the direction that light is traveling. We already know that light, like any wave, travels in a direction perpendicular to its planes of constant phase:

Figure 2.5.1 – Light Waves Travel in Several Directions at Once



So in our wave view of light, we say that the light wave is traveling in many directions at once, but now we are going to change our perspective to that of an observer and a source. When we do that, we narrow down all the possible directions of the light wave motion to a single line, which we call a *light ray*. This is a directed line that originates at the source of light, and ends at the observer of the light:

Figure 2.5.2 – Source and Observer Define a Ray



Alert

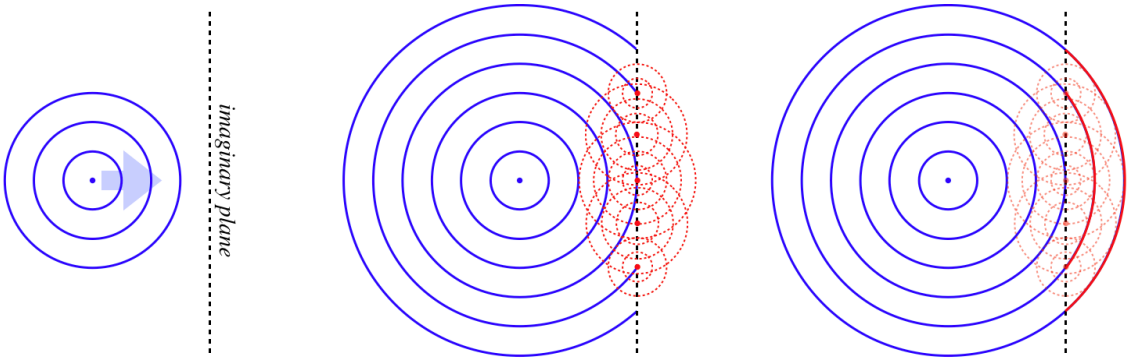
When most people encounter the idea of a light ray for the first time, what they think of is a thinly-confined laser beam. This is ***not*** *what is meant here! The ray has no physical meaning in terms of the confinement of light – we just use it as a simple geometrical device to link a source to an observer. Always keep in mind that the actual physical manifestation of the light is a wave that is usually traveling in many directions at once! Our use of rays will become so ubiquitous that this will be easy to forget.*

Reflection

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

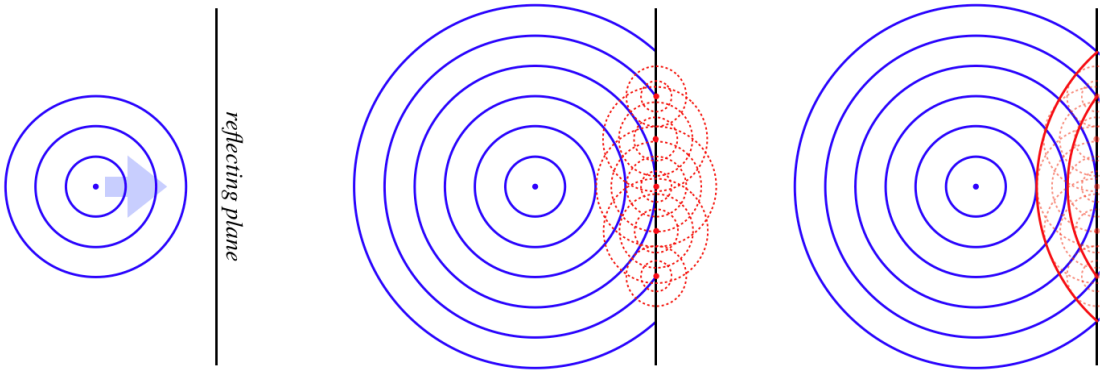
distances from the source to the plane, and when they are superposed, we know the result is what we see, which is a continued spherical wave (right diagram below).

Figure 2.5.3 – Spherical Wave Passes Through Imaginary Plane



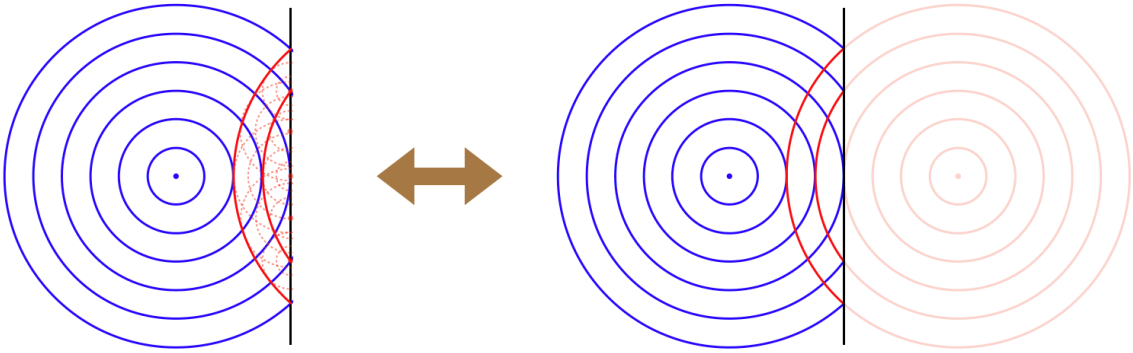
Now suppose the plane is not imaginary, but instead reflects the wave. Every point on this plane becomes a source of a wavelet, but this time, the wave created by these wavelets is going in the opposite direction. The wavelets have the same relative phases as in the previous case, and they are completely 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 2.5.4 – Spherical Wave Reflects Off 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:

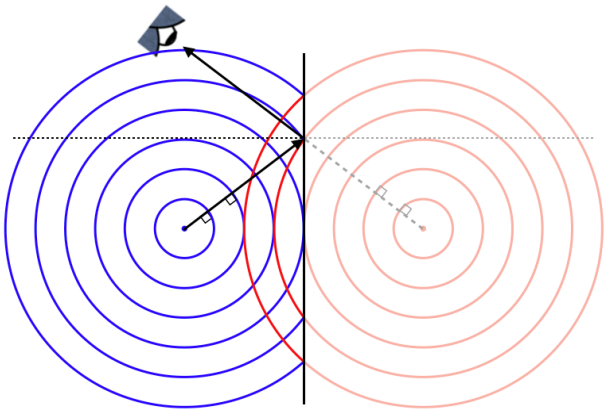
Figure 2.5.5 – Image of a Reflected Wave



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:

Figure 2.5.6 – Law of Reflection



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. This gives us the *law of reflection*, which states that the incoming angle (*angle of incidence*) equals the outgoing angle (*angle of reflection*):

$$\theta_i = \theta_r \tag{2.5.1}$$

The beauty of introducing rays is that from this point on, we can discuss sources and observers without a complicated reference to the spherical waves and Huygens's principle – we can just use the law of reflection and pure geometry.

Refraction

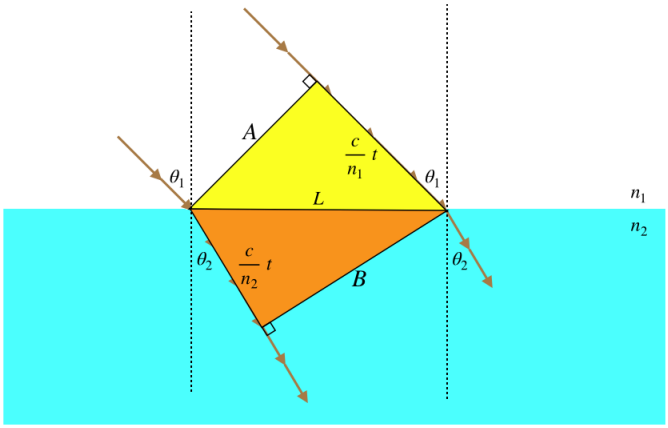
We saw that light waves have the capability of changing the direction of the rays associated with it through diffraction. We now consider another way that such a direction change can occur. This process, called *refraction*, comes about when a wave moves into a new medium. To get to the essence of this 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 2.1.2 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 3.6.7 – 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. In other words, it depends upon the indices of refraction of the two media. We can actually calculate this effect by freezing the figure above and looking at some triangles:

Figure 3.6.8 – The Geometry of Refraction



We are looking at what happens to a wavefront when it passes from position A to position B. The left side of the wave front is traveling within medium #2, during the same time period that the right side is traveling through medium #1. The rays are by definition perpendicular to the wavefronts, and we have defined the angles the rays make with the perpendicular in each medium as θ_1 and θ_2. Before we do any of the math at all, we immediately note:

Light passing from a faster medium into a slower medium bends toward the perpendicular, and light passing from a slower medium to a faster medium bends away from the perpendicular.

While the second of these conclusions is not expressed in our figure, it's not hard to see that it must be true, if we just imagine the wavefronts in the figure moving up to the left from medium #2 to medium #1.

Now for the math. We have two right triangles (yellow and orange) with a common hypotenuse of length we have called L. The distance between wavefronts in the upper medium is the speed of the wave there ($\frac{c}{n_1}$) multiplied by the time spent propagating, while the distance measured within the lower medium is calculated the same way, with a different speed ($\frac{c}{n_2}$). The angle θ_1 (shown on the right side of the diagram) is clearly the complement of the acute angle on the right-hand-side of the yellow triangle, which makes it equal to the acute angle on the left-hand-side of the yellow triangle. We therefore have:

$$\sin\theta_1 = \frac{\left(\frac{c}{n_1}\right)t}{L} \tag{2.5.2}$$

Similarly we find for θ_2:

$$\sin\theta_2 = \frac{\left(\frac{c}{n_2}\right)t}{L} \tag{2.5.3}$$

Dividing these two equations results in c and L dropping out, leaving:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{2.5.4}$$

This relationship between the rays of a light wave which changes media is called *the law of refraction*, or *Snell's law*. While this works in either direction of light propagation, for reasons that will be clear next, it is generally accepted that the "1" subscript applies to the medium where the light is coming from, and the "2" subscript the medium that the light is going into.

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:

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1 = 2.0 \cdot \sin 45^o = 1.4 \tag{2.5.5}$$

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:

$$n_1 \sin\theta_c = n_2 \sin 90^o \quad \Rightarrow \quad \theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{2.5.6}$$

Figure 2.5.9 – Partial and Total Internal Reflections By Incident Angle

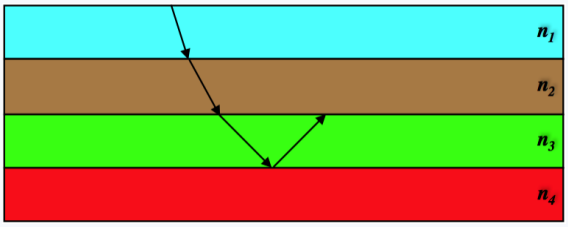
Note that there is at least partial reflection (obeying the law of reflection) every time the light hits the surface, but all of the light along that ray is only reflected when the ray's angle exceeds the critical angle.

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 2.5.1

The diagram to the right 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 $n_3 > n_4$.

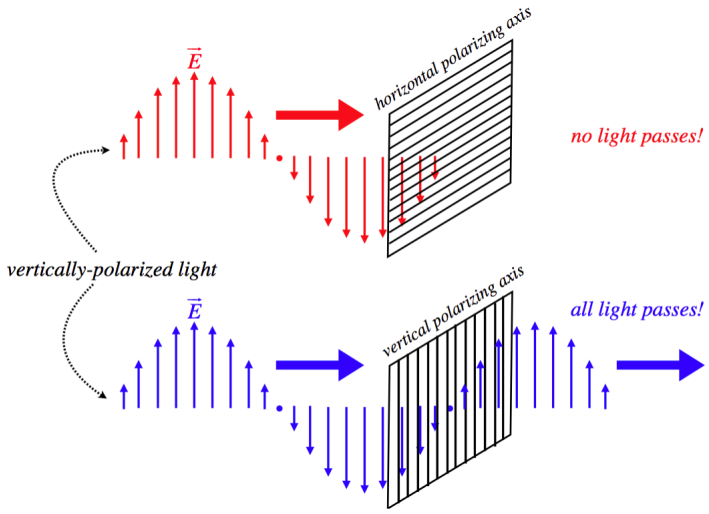
2.6: Polarization

Polarization Filters (Polaroids)

As stated previously when discussing the speed of light waves through transparent media, the mechanisms that govern light propagation through media are complicated. There is little we can say about it in this class, except to say that because the light wave is electromagnetic in nature, it interacts with electric charge, which is present in all matter. It so happens that it is possible to construct a solid substance which greatly restricts oscillatory motion of electric charges along a single dimension. The upshot of this is that the charges react to electric fields along one direction (or rather, components of electric fields along one direction), while they don't react along a perpendicular direction.

This material can have a dramatic effect on light passing through it. If the light is plane-polarized (see Figure 2.1.1), then its propagation through a medium will be affected by the preferential orientation of charge oscillations. When the light polarization is aligned with what we define as the *polarizing axis* of the substance, then little of the light is absorbed by the substance (i.e. the substance is transparent to this light), while if the light is polarized perpendicular to the polarizing axis, then virtually all of the light is absorbed. Such a filter is called a *polaroid* or *polarizer*.

Figure 2.6.1 – Light Through a Polaroid



One interesting application of this phenomenon is 3-D movies. Long ago someone came up with a brilliant idea for making movies projected onto a 2-D screen appear in 3-D. The idea is based on the fact that a large component (but not the only one) of seeing in 3-D is stereo vision. Your right eye sees objects from one perspective, while your left eye sees it from a slightly different perspective. You can see this is true by holding up your finger in a fixed position and alternately opening-and-closing each eye. Your finger's position appears to change relative to the background. This inventor's idea was to project not one but *two* images on the same screen. One image is recorded from the perspective of the right eye, and the other from the perspective of the left eye. Then the trick is to make the right-perspective image invisible to the left eye and the left-perspective image invisible to the right eye, so that each eye sees only its own perspective. The original inventor did this with colors – red lenses obscure red images, and yellow lenses obscure yellow light, so films were recorded from two perspectives, and each perspective was projected in a different color – one red and one yellow. But today we like our movies to be in realistic colors, so someone came up with the idea of projecting the two images with differently-polarized light, and then give viewers glasses that only admit the properly-polarized light into the respective eyes.

We have overly-simplified things here, in a couple of ways. First of all, a light wave does not have to arrive at the polarizer in either a parallel or perpendicular orientation – it could be aligned at any angle with the polarizing axis. What happens then? Well, electric fields are *vector fields*, which means they can be broken into components, so the component of the electric field that is parallel to the polarizing axis gets through, and the other component is absorbed.

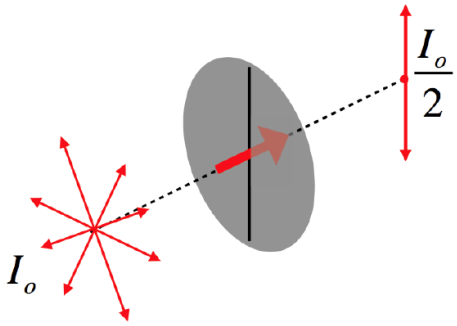
The second oversimplification is that not all of the individual light waves that come from a source are necessarily polarized in the same direction. In fact "natural" light from light bulbs and the sun is "unpolarized," which comes about because each of the individual light sources (atoms) are aligned in random orientations, and all send out random, unaligned light waves. When such

light is passed through a polaroid, half the light gets through. To see why this should be so, break every electric field vector of every wave into components parallel and perpendicular to the polarizing axis. Because the wave polarization directions are randomly-oriented, there is no reason to expect there to be a greater sum of components along one axis than another. By "half the light gets through," what do we mean? We mean that the intensity drops by one half. We look at the more general case of intensity next.

Intensity

We can express the fact that half of natural light gets through a polaroid in a diagram as follows:

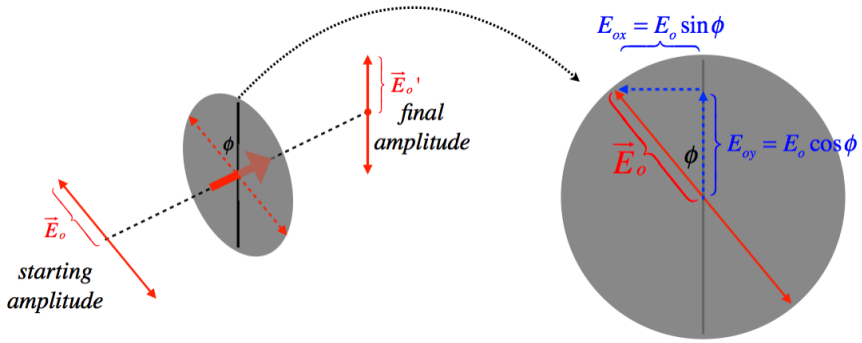
Figure 2.6.2 – Natural Light Intensity Halves Through a Polaroid



Now let's consider what happens if we send the natural light through *two* polaroids in succession. Clearly when the light reaches the second polaroid it will be plane-polarized from the first one. If the second polaroid is oriented the same as the first, then all the light gets through, and the intensity is unchanged, and if its polarizing axis is at right angles to the first polaroid, then no light will get through it. But now we seek to determine the intensity of the light that passes through the second polaroid if the angle between their polarizing axes is somewhere between 0^o and 90^o.

This process all comes down to what happens to the electric field vectors. After passing through the first polaroid, all the electric field vectors are aligned with that polaroid's polarizing axis. When those vectors come upon the second polaroid, just the component of the field vector that is aligned with the new axis gets through, resulting in a new vector shorter than the original.

Figure 2.6.3 – Polarized Light Through a Polaroid



Resolving the original electric field vector into components parallel and perpendicular to the polarizing axis, and keeping only the parallel part means that the new electric field vector magnitude is:

$$E_o' = E_o \cos\phi \tag{2.6.1}$$

The electric field vector is the *amplitude* of the light wave, and we are interested in the *intensity*. As with any other wave, the intensity is proportional to the square of the amplitude, so the relationship between the outgoing intensity I and incoming intensity I_o is:

$$I = I_o \cos^2\phi \tag{2.6.2}$$

This is known as *Malus's law*. Notice that it works exactly as we expect for the cases where the angle happens to be 0^o and 90^o.

Example 2.6.1

Unpolarized light enters a series of four polaroids with axes of polarization that are each rotated 30^o clockwise from the previous polaroid, making angles of 0^o, 30^o, 60^o, and 90^o with some common reference point. What fraction of the intensity of the incoming light is the intensity of the outgoing light?

Solution

When the unpolarized light passes through the first filter, the intensity is cut in half and comes out polarized at 0^o. Then it passes through three successive filters, and applying Malus's law for each 30^o change of polarization angle brings in a factor of 0.75 for each polaroid. The result is that the final intensity is:

$$I = I_o \left(\frac{1}{2}\right)\left(\cos^2 30^o\right)^3 = I_o \left(\frac{1}{2}\right)\left(\frac{3}{4}\right)^3 = \frac{27}{128} I_o$$

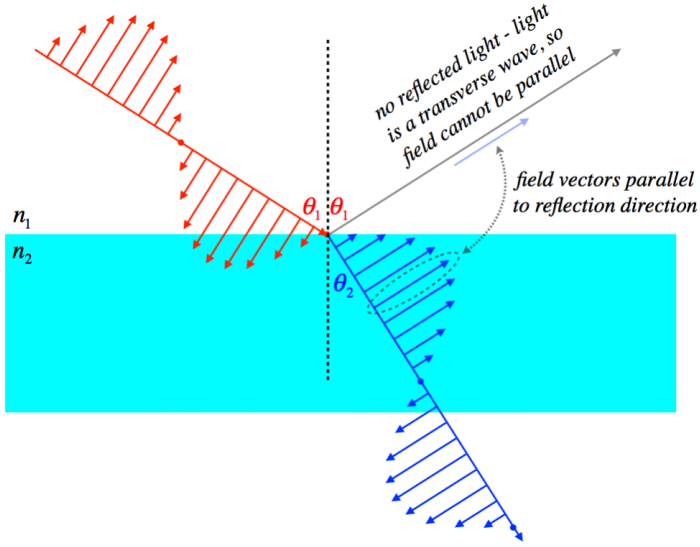
*One might expect that since the first and last polaroids are at right angles to each other, no light at all should emerge from the last polaroid. But when the light passes through a polaroid, it gains a new polarization aligned with that polaroid's polarization axis, and has no "memory" of its previous plane of polarization. Unless two **consecutive** polaroids are at right angles, some light will always get through each polaroid.*

Polarization By Reflection

While most natural light is unpolarized and we can polarize it with a polaroid, it turns out that is not the only way it can be polarized. A more "natural" way to create polarized light exists thanks to reflection. As we have said many times, when light (or any wave) strikes an interface between two media, it is partially transmitted and partially reflected.

Consider the following scenario: Light polarized in the vertical direction strikes an interface between media such that the reflected ray aligns with the electric field vectors of the transmitted ray. There is an important principle in physics that states that the conditions at the boundary have to work out properly. This means that the electric field vector of the incoming light must add up properly to the electric field after striking the interface. The electric field vector can of course be written in components with the "x-direction" being the electric field direction of the transmitted wave, and the "y-direction" being the direction of the reflected ray (which is perpendicular to the transmitted ray). But the outgoing light cannot have an electric field vector pointing along its direction of motion (light is a transverse wave), so no light reflects!

Figure 3.7.4 – Reflection of Polarized Light



Perhaps a more physically-intuitive and satisfying (if less mathematically rigorous) explanation for this effect involves the role of the electric charges. The electric field in the EM wave exerts forces on electric charges in the material from which the wave is reflecting, and these forces cause the charges to oscillate along those electric field directions. The oscillating charges in turn produce their own EM wave, which propagates in a direction perpendicular to their oscillations. At the angle shown above, the charges in medium #2 are oscillating in a direction parallel to the direction that a reflected wave needs to go according to the law of reflection, but charges cannot produce light waves parallel to their direction of oscillation, so no wave is reflected.

Of course this result is only for vertically-polarized incoming light, so unpolarized light that reflects at this angle will have its vertical component removed, which means that the reflected light is horizontally-polarized. More generally, light that is reflected off a surface at just the right angle will be polarized *parallel to that surface*. It also happens that if the angle is not just right, then while the light is not entirely polarized, it is partially so (depending upon how close to the correct angle the reflection is). By "partially polarized," we mean that the amplitude of light waves measured (using a polaroid) along one direction is not the same as the amplitude measured along the orthogonal direction. In practice this means that a polaroid aligned parallel to a surface from which the light is reflected will admit more light than a polaroid aligned perpendicular to that surface.

We can easily write down an expression for the "special angle" at which total polarization occurs (this is known as the *Brewster angle*), by noting that for this angle the reflected ray makes a right angle with the transmitted ray (because the field vector of the transmitted wave is perpendicular to the transmitted ray and is parallel to the reflected ray). Combining this fact with Snell's law gives the Brewster angle, θ_B:

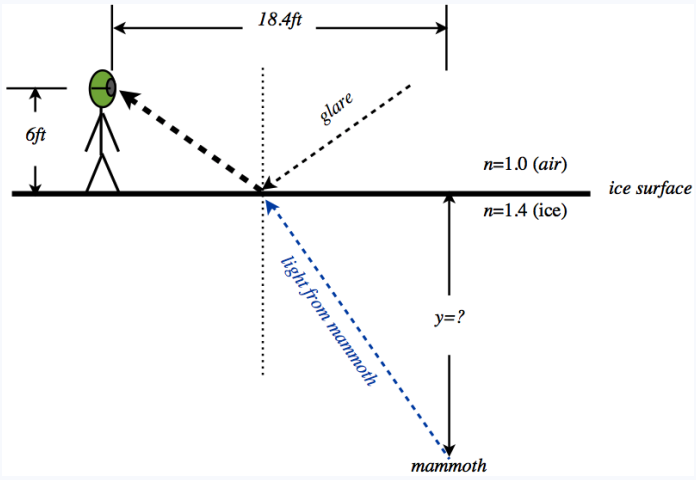
$$n_1 \sin\theta_B = n_2 \sin\theta_2 = n_2 \sin(90^o - \theta_B) = n_2 \cos\theta_B \quad \Rightarrow \quad \tan\theta_B = \frac{n_2}{n_1}\,, \tag{2.6.3}$$

where n_1 is the index of refraction of the medium within which the reflection is occurring, and n_2 is the index of refraction of the medium off which the reflection is occurring.

A nice application of this effect involves polaroid sunglasses. Most glare from sunlight comes off surfaces that are horizontal (roads, lake surfaces, etc.), which means that the light that reflects off such surfaces has a relatively small fraction of its polarization in the vertical direction. This means that if we place polaroids in front of our eyes that are allow only vertically-polarized light to pass, then very little of the horizontally-polarized glare gets through. Of course, only half of the non-glare light gets through as well, but at least one's vision of light of important objects (on coming cars or boats, etc.) does not have to compete with the incoming light from glare.

Example 2.6.2

A paleontologist is looking for the remains of a wooly mammoth in an unusually clear section of a glacier. The glare off the ice from the sun makes it hard for her to see, so she puts on her polarized sun glasses and is immediately rewarded when, along the line where the glare is cut to zero, she finds what she is looking for. Now she just needs to figure out how deep the carcass is. Fortunately she has a physicist (you) on staff. You measure the height of her eyes above the ice surface to be $6\ ft$*, and you measure the distance from the position where she first saw the beast through the glare, to the point where you can look straight down at it. This distance is* $18.4\ ft$*. You estimate the index of refraction of the ice to be 1.4. Find the depth of the wooly mammoth.*



Solution

For the polarized sunglasses to remove all the glare, the angle the light makes with the perpendicular to the ice must be Brewster's angle, so:

$$\tan\theta = \frac{n_2}{n_1} = 1.4$$

From the right triangle on the left, we can derive the distance from the paleontologist to the point of reflection:

$$\tan\theta = \frac{x_1}{6\ ft} \quad \Rightarrow \quad x_1 = (1.4)\,(6\ ft) = 8.4\ ft$$

We can use this distance to derive the horizontal distance from the point of reflection to the point on the ice directly above the mammoth:

$$x_2 = 18.4\ ft - 8.4\ ft = 10.0\ ft$$

The Brewster angle occurs when the reflected light makes a right angle with the transmitted light, and from symmetry (just reverse the direction of the light to see this), that is also true of the incoming glare and the light from the mammoth. Therefore we can use x_2 and the tangent of the angle to get the depth:

$$\tan\theta = \frac{y}{x_2} \quad \Rightarrow \quad y = (1.4)\,(10.0\ ft) = 14.0 ft$$

CHAPTER OVERVIEW

3: "Wait, what?" Experiments Reveal Cracks in Our Understanding

---

3.1: Blackbody Radiation

"Thermalizing" Energy

Two of the most important achievements in theoretical physics happened within only a very short span of time, in the mid-to-late 19th century. One of these was by Ludwig Boltzmann, who provided a description of thermal physics (and particularly the entropy function) in terms of a statistical model. [We made use of his advances in Physics 9HB.] The other achievement was perhaps even more profound, and occurred roughly a decade before Boltzmann's description of entropy in terms of microstate multiplicity. James Clerk Maxwell (who also had a hand in describing thermal/statistical physics, and shares credit for the "Maxwell-Boltzmann distribution") mathematically unified the electric and magnetic forces, thereby providing an electromagnetic explanation for the phenomenon of light. It's only natural that work that followed the contributions of these two giants at the end of the 19th century would veer toward the interplay between these two ideas.

It doesn't take modern science to understand that the same light that allows us to see can also make materials hot when it is absorbed. It also comes as no surprise that when objects get very hot, they can give off light. Indeed, this is how Edison's light bulb worked – the filament gets very hot, and out comes the light! And since we know that light comes in a very wide spectrum that goes well beyond the visible in both directions, we can conclude that objects of all temperatures emit EM radiation, and that the type or amount of radiation must somehow depend upon the object's temperature. This question of the relationship between an object's temperature and the light it radiates falls squarely into this cross-over of Boltzmann's and Maxwell's work.

At the core of this question are two facts, one from each of the two fields of physics:

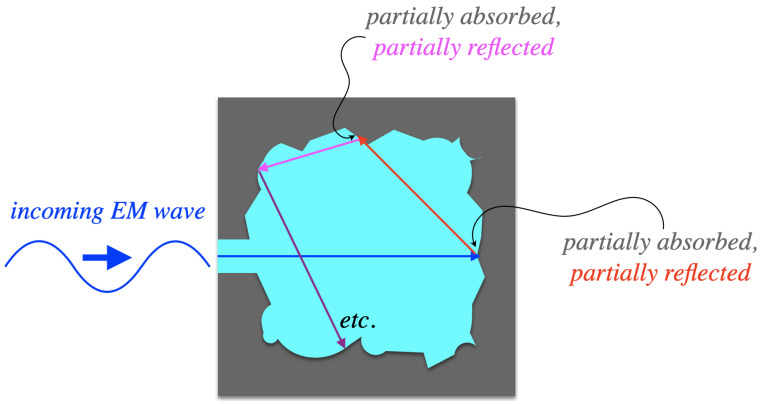
- The energy contained in the random movements of tiny particles (in our case, we are mostly interested in electrons), is what we refer to as "thermal energy", and according to kinetic theory, temperature and thermal energy are more-or-less proportional values.
- These same randomly-moving particles often have electric charge, and according to electromagnetic theory, accelerated electric charges are the source of EM radiation.

The interesting point here is that *randomly* accelerated charged particles will vibrate at a large variety of frequencies, and since these vibrational frequencies match the frequencies of the light waves they emit, the emission of a hot object must have a variety of frequencies as well. So two questions now come to mind:

1. How does the rate of energy emission from a hot object (in the form of radiation), relate to the temperature of that object?
2. How is the energy that is emitted from a hot object distributed across the spectrum of frequencies of the emitted waves?

To answer these questions experimentally requires some clever steps be taken to maintain controls on the experiment. For example, if we take careful measurements of light coming from a hot object, might not some of that light simply be light that originated outside the object, and is reflected off it? If the object is blue in color, then the energy that comes from the object in the reflected blue light throws-off the calculation for the radiated power in the blue region of the spectrum due to the thermal properties alone. So ideally, what we want to measure the radiate from is a *blackbody*. This is an object that effectively doesn't reflect any light. Any light that goes into it can have any distribution of frequencies whatsoever (it can even be monochromatic), but it is totally absorbed by the blackbody, with its energy "thermalized" into random particle motions within the body. With such an object, one can measure the radiation emitted, safe in the knowledge that all of it is from a thermal source of a known temperature. How do we create (or rather approximate) such a creature?

Figure 3.1.1 – A Blackbody Cavity



Suppose we create a cavity with irregular interior walls, and a small hole that connects the interior and exterior. Light will be reflected off the outer surface of this object, but in the region of the hole, light that goes in gets reflected around enough times that it is effectively entirely absorbed, and doesn't reemerge from the hole. This means that all the radiation that comes out of the hole can only have the thermal vibrations of the particles inside the cavity as a source – blackbody radiation! We can now stick a thermometer into the cavity, and take measurements of the light that emerges to answer questions 1 and 2 above. It should be noted that if the object is at the same temperature as its surroundings, then it is in thermal equilibrium with its environment, which means that for every joule of energy that enters the hole, one joule also emerges from it. But the light that goes into the cavity can be of any mix of frequencies, while the light that emerges must come out in a specific distribution across the spectrum that is a function of only the temperature.

Question 1 above doesn't consider at all the distribution across the various frequencies of emitted light, so it is the easier of the two questions to understand and answer. This question was answered by Josef Stefan in 1877, by analyzing data from an experiment performed by John Tyndall 13 years earlier. Stefan's empirical answer was later confirmed theoretically by a calculation performed by Boltzmann (which we will not reproduce here), yielding what is now known as the *Stefan-Boltzmann law*:

$$P_{blackbody} = \sigma A T^4 \,, \quad \sigma = 5.67 \times 10^{-8} \frac{W}{m^2 K^4} \tag{3.1.1}$$

The "P" in this equation is power, and the "A" is the surface area of the blackbody (from which the radiation emanates). For our little model above, this would be the area of the tiny hole, out of which the energy radiates. For what we generally think of as a "blackbody" which radiates in every direction, it is the full surface area of the object. The only adjustment to this law involves an additional factor e (called the *emissivity*), which is a number less than or equal to 1 that takes into account the fact that the object might not be a perfect blackbody ($e = 1$).

This result for the total emission of the blackbody is not what interests us most here. It is question 2 that starts us down the crazy path on which we are about to embark.

Rayleigh-Jeans Law

The cavity version of a blackbody gives us another picture that is helpful for analysis, if we ask what is going on *inside* the cavity at steady-state. Clearly EM waves are being bounced around inside, and occasionally a wave has just the right direction to head out of the hole, while occasionally a wave enters the hole as well, keeping the energy inside the cavity constant. Okay, so this language makes it sound like the emergence of a wave from the hole is a rare event, when in fact there are constantly light waves passing in and out, but the point is that inside the cavity is a flurry of light wave activity. Each of these waves carries energy, and as a finite amount of energy is distributed among a random sampling of waves, we can use what we learned in 9HB about the probability of finding that a single entity in a collection has a specific energy. We found that according to the Boltzmann distribution, the average energy per particle in a system at temperature T is:

$$\langle E \rangle = \frac{\sum\limits_{\text{all E}} E\, e^{-\frac{E}{k_B T}}}{\sum\limits_{\text{all E}} e^{-\frac{E}{k_B T}}} \tag{3.1.2}$$

[Reminder: $k_B = 1.38 \times 10^{-23}\ J/K$ is the *Boltzmann constant*.]

From what we have learned about waves, we know that their energy is a function of their characteristics. For example, we found that the energy in a single wavelength of a wave on a string (Equation 1.3.7) is proportional to the frequency and the square of the amplitude (as well as some characteristics of the medium, such as the string density and the speed of the wave). So given the continuum of frequencies and amplitudes possible, the possible energies lie on a continuum, and our "sums" are really integrals:

$$\langle E \rangle = \frac{\int_0^\infty E\, e^{-\frac{E}{k_B T}}\, dE}{\int_0^\infty e^{-\frac{E}{k_B T}}\, dE} \tag{3.1.3}$$

These are not difficult integrals to perform (or look up), and the end result is that the average energy per particle should be:

$$\langle E \rangle = k_B T \tag{3.1.4}$$

This seems reasonable, but next we need to see what this implies for the intensity of the EM waves coming from a blackbody *as a function of their frequencies*. Let's consider what happens if we look only at light of a single frequency. This subset of all the waves still come in a random distribution, but now the only characteristic of the waves that is random is the amplitude. The average energy of each of these waves is fixed at $k_B T$, so the total energy of this group of waves with this specific frequency is just the number of waves in the group multiplied by this average. If we look at group characterized by another frequency, once again the waves in the group have the same average energy, but there may be more or fewer than the previous group. So for example, if we see twice the number of waves at frequency f_1 as we see at f_2. With each wave getting the same average energy of $k_B T$, this would mean that the light that emerges with frequency f_1 would be twice as intense as the light that emerges with frequency f_2.

Of course, we can't really talk about the number of waves at a *precise* frequency, since frequencies lie on a continuum (two randomly-chosen waves could never have exactly the same frequency). Instead, we can only talk about the number of waves that lie within a *range* of frequencies. So for example, we could compare the number of waves we see with a range of frequencies between f_1 and $f_1 + \Delta f$ to the number we see in the same-sized frequency range between f_2 and $f_2 + \Delta f$. Then the question becomes, if we take two ranges of frequencies, one of them near f_1 and the other near f_2 (and let's say that $f_2 > f_1$), in which of these ranges will we find more individual waves? We need to know this, because the amount of energy in that range (which directly relates to the light intensity in that range) is the number of waves multiplied by the average energy per wave.

If this was a 1-dimensional wave, then the number of waves in each range would be equal, because such waves are uniquely-defined by their frequencies. But in three-dimensions, the wave has more degrees of freedom, and many different waves can have the same frequency. It turns out that the higher the frequency, the more such waves are possible. The rate at which the number of waves grows with respect to the increase in frequency is called the *density of states*. All densities are an amount of something *per* something else (like mass density is mass *per* volume), and this is the number of "states" one can find a wave in *per* frequency. We will not go into a derivation of this quantity for a collection of light waves with frequency f at equilibrium in a confined space of volume V (this is referred to as a *photon gas*), but it comes out to be:

$$\frac{dN}{df} = \frac{8\pi V}{c^3} f^2 \tag{3.1.5}$$

Here one can think of V as the volume of the blackbody cavity (volume occupied by the photon gas), and of course c is the speed of light. But more generally, V can be any arbitrary volume where this radiation is present.

Okay, so returning to our original question of the intensity of blackbody radiation as a function of frequency of that radiation, we can now say how much energy there is collectively in the waves within a frequency range from f to $f + df$. The energy in that range is the energy per wave times the number of waves. In this infinitesimal range, of course the energy is infinitesimal:

$$dE = \langle E_{wave} \rangle \cdot dN = \frac{8\pi k_B T}{c^3} V f^2 df \tag{3.1.6}$$

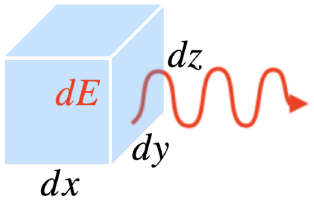
It's generally easier to talk about an energy density, as it is defined at specific points in space and does not depend upon defining a volume. Dividing the energy by the volume, we get:

$$u \equiv \frac{E}{V} \quad \Rightarrow \quad \Psi(f,T) \equiv \frac{du}{df} = \frac{8\pi k_B T}{c^3} f^2 \tag{3.1.7}$$

The function obtained after dividing through by df indicates how the energy density at a point in space relates to the range of frequencies of EM waves at that point. That is, integrating the function $\frac{du}{df}$ over a range of frequencies gives the energy density that results from the EM waves that are in that frequency range. This function therefore not only involves a density in space, but also over frequencies. This latter density in frequency gives it its name: *spectral energy density*. This result is known as the *Rayleigh-Jeans law*.

The reader may be concerned that we seem to have used "energy density" and "intensity" interchangeably – as if the energy density of the light at a point in space is synonymous with its brightness. In fact these two quantities are very closely related to each other. Consider a small cubical region in space, through which a plane light wave is moving parallel to two sides. This energy is being carried by a light wave, and all of it passes out of the cube, through one of the cube's faces.

Figure 3.1.2 – Relating Energy Density to Intensity



The time it takes this to occur is the distance traveled (the length of one edge of the cube), divided by the wave speed:

$$dt = \frac{dx}{c} \tag{3.1.8}$$

The power delivered through the cube face is the energy divided by this time:

$$P = \frac{dE}{dt} = \frac{dE}{dx}c \tag{3.1.9}$$

The intensity is the power delivered divided by the area through which it passes:

$$I = \frac{P}{A} = \frac{\frac{dE}{dx}c}{dydz} = \frac{dE}{dxdydz}c \tag{3.1.10}$$

And the energy density in this region is the total energy divided by the volume, $dV = dxdydz$, so the intensity and energy density only differ by a factor of c.

While this is true for a plane wave, for our light coming from the tiny hole in our blackbody cavity, the light spreads as it exits. This change of geometry leads to a smaller intensity (imagine the same light wave in the diagram above coming out of five of the cube's surfaces, rather than just one). It's far to much detail to go into the effect this has on the case of the relation between the energy density and the intensity of the radiation coming from our blackbody (the reader can look up "Lambert's cosine law", if the desire to fall down a deep rabbit hole is desired), but the upshot is that it changes the energy density/intensity relation by an additional multiplicative factor.

Wien's Contribution

Let's take a moment to ask the most important type of question in physics: "What does the Rayleigh-Jeans result predict, physically?" Suppose we filter the light coming from a blackbody according to frequency (for the visible spectrum, we simply could do this with a diffraction grating), and measure the intensity of the light as the frequency rises. We should see that the higher frequencies are always brighter than the lower ones, since the energy in a small range grows indefinitely in proportion to f^2. In the case of visible light, we would see orange brighter than red, yellow brighter than orange, and so on, forever, *including* frequencies above the visible spectrum. There is only a finite amount of energy available, and clearly there is *some* energy at the lower frequencies that we measure, so since there is no upper-bound on the frequency the light can have, the total energy becomes unbounded. This disaster of a prediction became known as the "ultraviolet catastrophe", because for the temperatures being studied (namely the temperature of the surface of the sun, which approximates a blackbody and gives us ample light to observe), the energy content of the light actually starts to *decline* (not increase unbounded) around the high end of the visible spectrum.

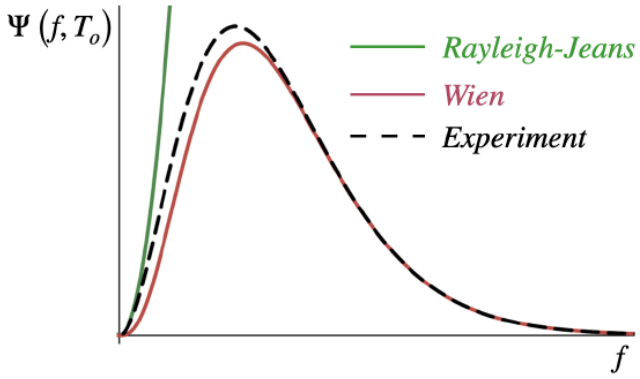
Despite the "catastrophic" result for frequencies beyond a certain frequency, the Rayleigh-Jeans result actually predicts results pretty well for low frequencies. But as indicated above, the intensity of the light peaks at some value of frequency, and comes down

again. This behavior, as well as a very good approximation to the observed results for high frequencies was modeled by a fellow named Wilhelm Wien. This approximated result Wien obtained was:

$$\Psi(f,T) = \frac{8\pi h}{c^3} f^3 e^{-\frac{hf}{k_B T}} \tag{3.1.11}$$

The constant h that appears here was not computed by Wien (it is included here for later comparison) – he was just proposing the general functional form. This function certainly looks significantly closer to the experimental results for this curve, but it suffered from two ailments that Rayleigh-Jeans did not: The physics to explain its origin was not clear, and it failed to properly model the low frequency emissions of blackbodies.

Figure 3.1.3 – Rayleigh-Jeans, Wien, and Experimental Results



There was, however, one result that came from Wien's model that holds equally true for what we see in nature. The graph above is for a specific blackbody temperature (which we called T_o in the graph). If we look at the curve for the same blackbody at another (let's say higher) temperature, what do we see? We can already make one guess, from two things that we know:

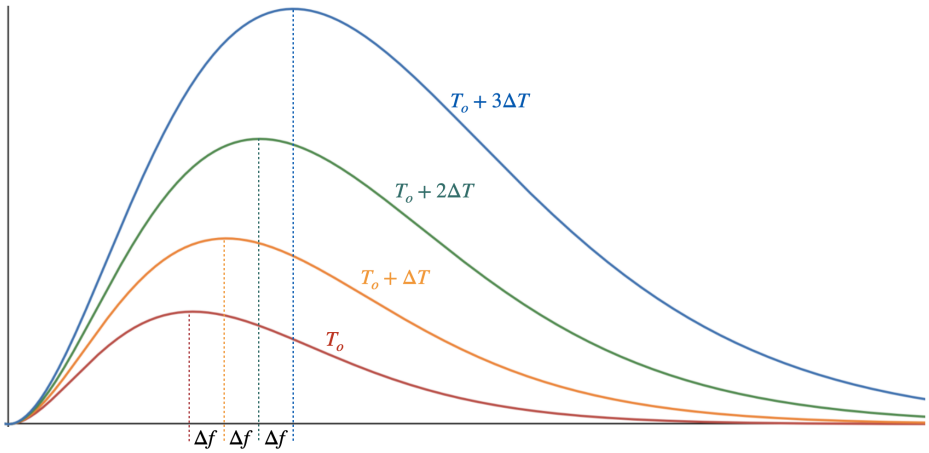
- This is a curve that represents energy density over frequencies. If we integrate a density function over a range, we get the total amount in that range. For example, if we integrate a mass density over a volume, we get the total mass within that volume. Therefore, if we want to know the light energy density (or equivalently, the light intensity) at a point in space that comes only from a range of frequencies, we integrate over the frequency range:

$$u(T) = \int_{f_1}^{f_2} \Psi(f,T)\, df \tag{3.1.12}$$

- The *total* power coming from the blackbody grows with the fourth power of the temperature, according to Equation 3.1.1. This total power is over the entire spectrum of frequencies.

Putting these together, it is clear that the area under the *entire* curve must grow as the temperature rises, and the curve has to maintain its general features, so we expect the peak value to rise. But it turns out that there is another effect that comes from changing the temperature as well: The peak of the curve displaces to the right with an increase in temperature, and to the left with a decrease. But more precisely, it displaces an amount that is *proportional to the temperature change*.

Figure 3.1.4 – Displacement of Blackbody Curve with Temperature



This (not coincidentally) is a feature of both Wien's curve and what is experimentally seen, though the constant of proportionality of the frequency where the peak occurs and the temperature of the blackbody is not the same in both cases. Using the correct constant, we have what is now known as *Wien's Displacement law*:

$$f_{peak} = \alpha T\,, \quad \alpha = 5.879 \times 10^{10}\,\frac{Hz}{K} \tag{3.1.13}$$

Planck's Puzzling Fix

Max Planck knew what the spectral energy density curve for a blackbody looked like, and decided that rather than try to figure out where the physics went wrong, he would see what math was required, and then try to work backwards to the physics. The calculation leading to the ultraviolet catastrophe came in two parts – the average energy per wave, and the density of states. There's really no physics in the latter calculation, so the only place where a change can be made is in the calculation of the average energy per wave. There was no arguing with Boltzmann's probabilities either, so Planck tried making a different assumption about the distribution of energy amongst the waves.

Looking back once again at Equation 1.3.7, we make a note that the energy of one wavelength of a string wave is proportional to its frequency and its amplitude-squared. If we limit ourselves to looking at the energy of only waves of a given frequency, then the energy of these waves depends only on the amplitude, but as Rayleigh and Jeans assumed, this amplitude is variable on a continuum. That is, if you wanted a light wave of frequency f to have slightly more energy with the same frequency, you could give it as little energy as you like, because you are free to increase the amplitude by an infinitesimal amount. Plank noticed that the calculation takes a different turn if we *don't* assume this. He speculated that perhaps there was a minimum amount that you could change the energy by, and that this amount was proportional to the frequency. That is, he posited that perhaps the possible energies for a light wave, instead of being continuous, might be discrete, and simply be a multiple of a minimum energy, ϵ:

$$\epsilon = hf \quad \Rightarrow \quad E_n = n\epsilon = nhf \tag{3.1.14}$$

The constant h, now known as *Planck's constant, is:*

$$h = 6.63 \times 10^{-34}\ J \cdot s \tag{3.1.15}$$

This assumption profoundly changes the calculation for the average energy per wave. With the energies no longer on a continuum, the Boltzmann distribution calculation of the average no longer becomes an integral, but instead Equation 3.1.2 becomes:

$$\langle E \rangle = \frac{\sum_{n=0}^{n=\infty} E_n\, e^{-\frac{E_n}{k_B T}}}{\sum_{n=0}^{n=\infty} e^{-\frac{E_n}{k_B T}}} = \frac{\sum_{n=0}^{n=\infty} nhf\, e^{-\frac{nhf}{k_B T}}}{\sum_{n=0}^{n=\infty} e^{-\frac{nhf}{k_B T}}} \tag{3.1.16}$$

These infinite series might seem daunting at first, but with a simple substitution, they take the appearance of common geometric series whose sums are well-known:

$$\alpha \equiv e^{-\frac{hf}{k_B T}} \quad \Rightarrow \quad \langle E \rangle = hf \frac{\sum_{n=0}^{n=\infty} n\alpha^n}{\sum_{n=0}^{n=\infty} \alpha^n} \tag{3.1.17}$$

These sums can be looked-up or computed (the denominator from a couple lines of algebra, the numerator by differentiating the denominator), and they are:

$$\sum_{n=0}^{n=\infty} n\alpha^n = \frac{\alpha}{(1-\alpha)^2} \qquad \sum_{n=0}^{n=\infty} \alpha^n = \frac{1}{1-\alpha} \tag{3.1.18}$$

Plugging back in for α and putting the series sums back in above gives:

$$\langle E \rangle = \frac{hfe^{-\frac{hf}{k_B T}}}{1-e^{-\frac{hf}{k_B T}}} = \frac{hf}{e^{\frac{hf}{k_B T}} - 1} \tag{3.1.19}$$

Multiplying this average energy by the density of states gives a very different spectral energy density than obtained by Rayleigh and Jeans, and differs from that produced by Wien only by a "-1" term in the denominator!

$$\Psi(f,T) = \frac{8\pi h}{c^3} \frac{f^3}{e^{\frac{hf}{k_B T}} - 1} \tag{3.1.20}$$

While this solution perfectly matched the experimental data, even Planck himself didn't believe the physics that leads up to it. He was certain someone would find an explanation for this curve other than assuming that the energy of a light wave is *quantized* into little packets that are multiples of hf.

3.2: The Photoelectric Effect

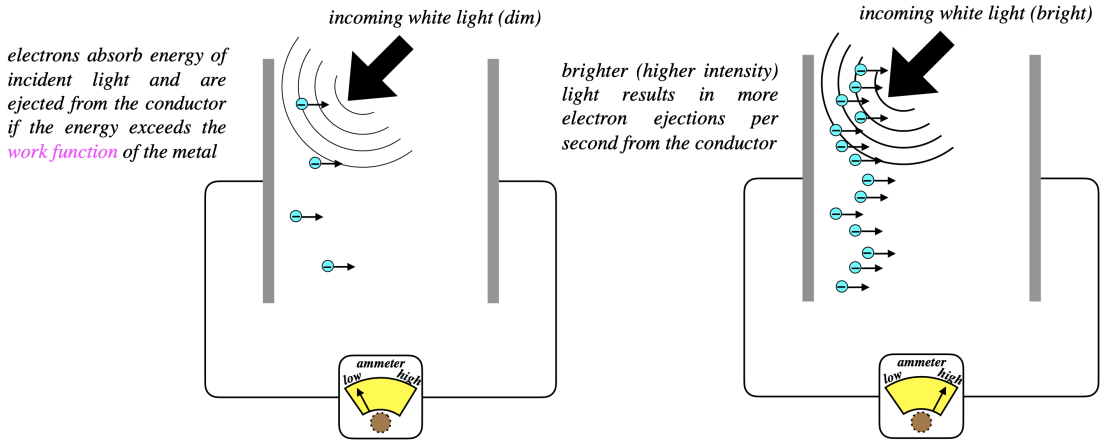
Light Interacting with Conductors

The common denominator of the problems that would plague physics for the early years of the 20th century involved light's interaction with matter. As the blackbody radiation puzzle showed, the simple view developed from Maxwell's EM and Boltzmann's thermodynamics were not sufficient to handle these problems. A whole new way of thinking about light and matter was needed. Planck started the revolution (without thinking it was correct), and the next bit of evidence would come from a 1905 paper by Einstein (yes, the same year as his paper on special relativity!), explaining a phenomenon relating to light striking a conducting surface.

In the study of classical EM, it is customary to assume that charges on the surfaces of conductors remained on those conductors. But this belies the fact that we sometimes see charges leap from a conducting surface (a spark) due to a strong external field. So we know that given enough additional energy (in the case of the spark, electrical potential energy), an electron *will* exit the surface of a metal (the protons are of course fixed within the lattice of the metal). Different metals will hold their electrons with differing degrees of "tightness," and this tightness is measured in terms of the minimum amount of energy needed to just barely free the most loosely-held electrons. This minimum energy for a given metal is called the metal's *work function*, typically represented by the symbol ϕ. Naturally an external static electric field is not the only way to give additional energy to these electrons, and it was known for quite some time that shining light on the metal can also add enough energy to the electrons to kick some off. When light accomplishes this, it is called the *photoelectric effect*.

At first glance, this phenomenon makes perfect sense – there is no sign of any of the "weirdness" that came out of Planck's explanation of the blackbody radiation curve a few years earlier. When light is shone onto the negative plate of a capacitor, some electrons are ejected and make their way to the positive plate. When the the missing electrons are replaced on the plate form the battery, the electron flow can be measured by an ammeter. If we turn up the brightness of the light, the measured current rises.

Figure 3.2.1 - Photoelectric Effect (Unsurprising)



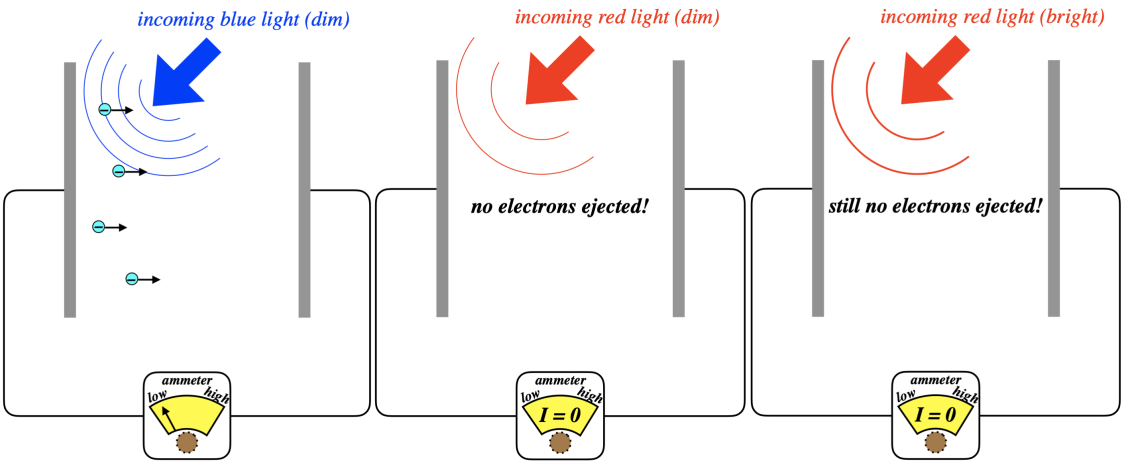
Digging Deeper

Physics is uninteresting if we are never surprised, so let's dig a little deeper and determine two other pieces of information, namely:

- Does this effect have any frequency dependence?
- What determines how much kinetic energy the electrons have after they exit the conductor?

So rather than just use white light, let's compare some monochromatic cases.

Figure 3.2.2 - Frequency Dependence of Photoelectric Effect



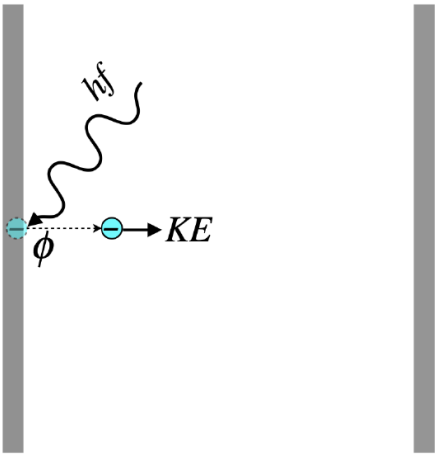
The blue light ejects electrons even when dim, as the white light did, but dim red light does not. This tells us that it was the blue part of the spectrum that was ejecting the electrons when the white light was shone on the metal earlier. While this result is peculiar from our standard understanding of EM, we can simply look at it as a confirmation of Planck's result from blackbody radiation: Blue light carries more energy than red light, since it is providing enough energy for the electrons to overcome the work function. So in order to see the same effect with red light as we saw with blue light, we just need to crank up the intensity of the red light to make up for the energy deficiency, right? No, it turns out it doesn't work this way at all!

Einstein explained the phenomenon in the following way: Notwithstanding light's obvious wavelike nature, in this setting it behaves like a particle (which we now call a *photon*), inasmuch as it can only be absorbed by a single electron, and only one photon strikes an electron at a time. We can call this the "one per customer" rule. This photon has an energy equal to hf (just as Planck found), and it gives all this energy to the electron it strikes.

Notice how perfectly this explains what we see. Any given electron must receive an amount of energy greater than the work function in order to be set free, but the most it can receive is hf, and if f is too low, then it won't be enough. The light doesn't behave like a wave in this case, which could continuously and gradually add energy to the electron until it has enough, but rather like a particle, in an all-or-nothing fashion. Furthermore, the *intensity* of the light is simply determined by the number of photons arriving per second. If the photons have enough energy to kick off electrons, then greater intensity means more electrons will be kicked per second, but if the individual photons don't have enough energy to kick off electrons, then adding more of them will not have any effect – they cannot "double-up" on an electron – there's only one per customer. Furthermore, a particularly energetic (high frequency) photon cannot split its energy between two electrons and eject them both.

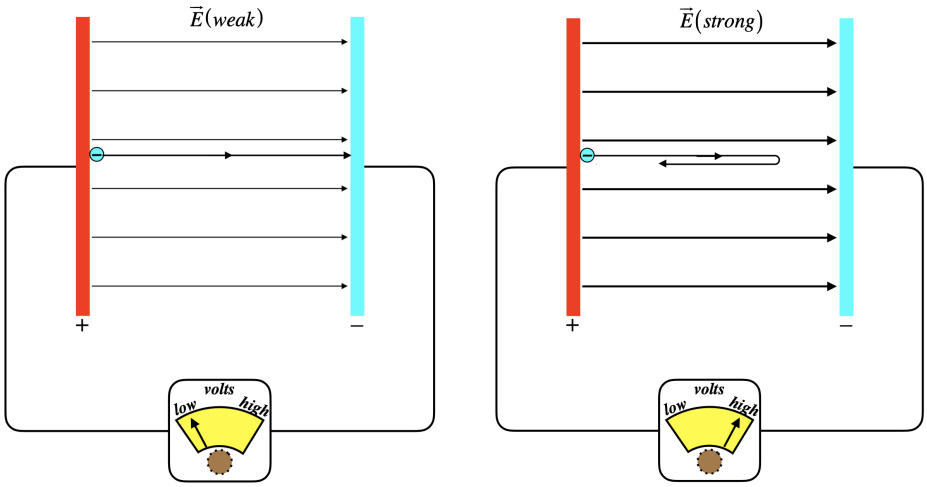
This answers the effect of frequency, but what about the second "digging deeper" question regarding the energy of the electrons that are ejected? Einstein's solution gives us that answer as well. Applying conservation of energy to this process gives us immediately what we seek:

Figure 3.2.3 - Photoelectric Effect Energy Accounting



From conservation of energy, we see immediately that of the energy introduced by the photon, some of it goes into the potential energy that is the work function of the metal (freeing the charge), and the remainder into the electron's kinetic energy. It should be noted that the work function is not a constant that applies to every electron – some will be bound more tightly to the metal than others. The work function is defined as the *minimum* binding energy for that metal – the energy required to tear away the easiest-to-remove electrons. This work function is found by measuring something called the *stopping potential*, which works like this:

Figure 3.2.4 - Stopping Potential



Here we are shining onto the positively-charged plate, ejecting electrons. The electrons come off the plate with some kinetic energy, but the electric field opposes their motion. If the field between the plates is weak, then some electrons will get across, and we can measure the flow. As we dial-up the strength of the field, however, fewer and fewer of the electrons will successfully make the journey. When the field is just barely strong enough to stop even the most energetically-ejected particles, then the potential energy that those electrons have to climb equals the kinetic energy at which they were ejected. As monochromatic light was used, every electron was given the same energy, so those that are ejected with the most kinetic energy are the ones held most weakly to the conductor. This minimum potential energy of the conductor is what we define to be its work function. Mathematically, the energy accounting looks like this:

$$e\Delta V_{stopping} = KE_{max} = hf - \phi \tag{3.2.1}$$

This equation is read this way: "The electron charge multiplied by the stopping (electrostatic) potential is the potential energy change that barely stops the electrons with the greatest amount of kinetic energy, and this equals the energy given to the electron by the photon, minus the work function (the potential energy holding the electron to the surface of the conductor)."

Since different electrons require different added energies to be torn away from the metal surface, this result is best described as a *definition* of the metal's work function. That is, the work function is the *weakest* "potential energy grip" that the metal has on all of its electrons:

$$\phi \equiv hf - KE_{max} \tag{3.2.2}$$

Applications

The applications of this effect are of course endless, as you can undoubtedly think of countless devices that involve detection of light. One interesting application is a device known as a *photomultiplier tube*. Suppose you wish to be able to detect and amplify very low intensities of light (in any part of the spectrum). Assuming you can find a metal with a low enough work function for the frequency of light you want to see, at low intensities the photons are only going to knock off a handful of electrons, which may not be particularly easy to detect. But the nice thing about converting a signal from photons to electrons is that we can add energy to electrons using electric fields, and electrons are also quite good (when propelled at sufficient *KE*) at knocking more electrons off a surface. Then those can do the same, and so on.

This device is indispensable for high-energy particle physics experimentation, when it is important to see where even a *single* photon produced in a certain collision lands. But it also works for common-use devices, such as night vision goggles. In this case, you have lots of photons landing in different places (i.e. an image focused by a lens), and each place where the photon lands has its

own tiny photomultiplier tube. Each tube constitutes one pixel, so all the tubes put together form an amplified image. This device is a step above an infrared sensing apparatus for applications that require better resolution of the image (we'll see why this is later), though it is constructed specifically for the visible spectrum, so it can't see through objects opaque to visible light, while some of those same objects may be (partly) transparent to infrared light.

3.3: Compton Scattering

Scattering

In Physics 9HA, we spend some time talking about collisions in two dimensions, such as two billiard balls. In Physics 9HB, we have learned that to treat these properly for fast-moving, tiny particles, we need to incorporate the corrected momentum from special relativity. We also learned in relativity that light carries momentum. This is not a property typically exhibited by waves, so it seems like studying momentum conservation in collisions between light and matter (called *scattering* of light off matter) might give us some insight into the new emerging idea from Planck and Einstein that light comes in individual packets with energy equal to hf.

We start with the energy/momentum of a photon first encounters in Equation 2.5.17 in Physics 9HB:

$$E_{photon} = pc \tag{3.3.1}$$

It can be shown that it is not possible for both energy and momentum to be conserved in the case of a free particle if the photon is absorbed, but if the photon scatters off the electron (i.e. a photon exits the collision), these conservation laws can be obeyed. A before/after picture of the event is helpful:

Figure 3.3.1 – Photon Scatters off a Particle

There is of course no reason to make this calculation more difficult than it needs to be, so choosing the "lab frame" (reference frame where the target particle is at rest) makes sense.

Just when a collision between a photon and an electron appears to be just like a collision between two billiard balls, the photon remembers that it is also a wave! If a cue ball collides with an eight ball, after it bounces off, it doesn't become a different ball, but a photon loses some of its energy (given to the electron), so since $E = hf$, its frequency goes down! In other words, the photon that leaves the collision is a totally different photon from the one that came in!

Compton Wavelength

Using the figure above, we can invoke momentum and energy conservation to relate the change in the photon's wavelength as a function of the scattering angle θ. Calling the momentum of the incoming photon p_{in}, the momentum of the outgoing momentum p_{out}, and using the relativistic momentum for the particle, we have the following two conserved momentum components:

$$\begin{aligned} x\text{-direction:} \quad & p_{in} = p_{out}\cos\theta + \gamma_u m u \cos\phi \\ y\text{-direction:} \quad & 0 = p_{out}\sin\theta - \gamma_u m u \sin\phi \end{aligned} \tag{3.3.2}$$

We can also have energy conservation at our disposal. The energy of the photons satisfy $E = pc$, and the particle initially has only it's rest energy, so:

$$p_{in}c + mc^2 = p_{out}c + \gamma_u mc^2 \tag{3.3.3}$$

After much algebra to eliminate the angle ϕ from the simultaneous equations, the result is:

$$\frac{1}{p_{out}} - \frac{1}{p_{in}} = \frac{1}{mc}(1 - \cos\theta) \tag{3.3.4}$$

We can take this a step further and compare the incoming and outgoing photon wavelengths, by using Planck's equation for the relationship between photon energy and frequency:

$$E_{photon} = pc = hf = h\left(\frac{c}{\lambda}\right) \quad \Rightarrow \quad p = \frac{h}{\lambda} \tag{3.3.5}$$

Putting this in above gives:

$$\lambda_{out} - \lambda_{in} = \frac{h}{mc}(1 - \cos\theta) \tag{3.3.6}$$

Let's take a moment to interpret what this means. We shine light of a known wavelength into a cloud of stationary particles (say electrons), and we measure the wavelengths of the light that come out at the various angles. From this data we can determine the mass of the particles in the cloud. The quantity $\frac{h}{mc}$ is often written as "λ_c", and is called the *Compton wavelength* of the particle with mass m. Notice that the most energy that the photon can lose is when it is *backscattered*, i.e. when it comes straight back the way it came in. In this case, $\cos\theta = \cos 180^o = -1$, which means that the wavelength of the incoming light is increased by two Compton wavelengths.

Another thing to note is that if the scattering is off a heavier particle (such as a proton rather than an electron), then the effect is far less pronounced, meaning the scattered light is closer in wavelength to that of the incoming light than if the particle were lighter. This makes sense, since heavier particles will take less of the photon's energy than lighter ones, so the outgoing photon energy will be closer to the incoming photon energy. A common way of stating this is to say that if the wavelength of the light is much greater than the Compton wavelength of the target particle, then the scattered light experiences a negligible wavelength shift compared to the incident light.

While we were not surprised my the fact that light carries momentum, because we already saw this when we studied special relativity, this Compton effect is a very clear example of a particle-like property possessed by light, as classical waves do not carry momentum. The evidence that light had particle properties just keeps mounting!

3.4: Matter Has Wave Properties, Too!

What About Known Particles?

Okay, so we have established that light can behave as a particle (photon) or as a wave (Maxwell). It seems as though how it behaves simply depends upon the context we put it into. If we do typical optics experiments (polaroids, double slit interference, etc.) then light is clearly a wave, but if we do the experiments of the 20th century (blackbody radiation curve, photoelectric effect, Compton scattering), then it acts very much like a particle. But how can a single entity simultaneously act like two such opposite phenomena? Is light localized into little packets of energy, or is it spread out and able to exhibit interference? IT CAN'T BE BOTH!!

Okay, let's set aside this conundrum for a moment and change gears. If light, which we previously "knew" to be a wave, can suddenly start showing particle properties, might we not see the wave properties for things we have always "known" to be particles? Well, in fact, just such an experiment has been performed by Davisson and Germer in 1927. They fired electrons through a lattice of nickel atoms, and found that the pattern they formed on the other side looked just like that of a diffraction grating. So not only do "known" waves have particle properties like momentum and localized energy, but "known" particles have wave properties like interference!

de Broglie's Idea

Unlike the case of Planck, who essentially worked backwards from puzzling experimental data, Louis de Broglie anticipated the results of Davisson and Germer, when a few years before their experiment, he hypothesized that if light can behave as both wave and particle, why not matter? But he did more than just say that this could be true, he found a common element between the two cases, and used it to make a prediction. It goes like this...

Light carries energy, and has momentum. According to what we learned in relativity, light has no mass, so these quantities are linked by:

$$E^2 = p^2c^2 + m^2c^4 \quad \Rightarrow \quad E = pc \tag{3.4.1}$$

But now we also know that the energy of a quantum of light is directly related to its frequency, so we can relate the momentum of light to its wavelength:

$$p = \frac{E}{c} = \frac{hf}{c} = \frac{h}{\lambda} \tag{3.4.2}$$

de Broglie speculated that perhaps the link between momentum and wavelength is a fundamental one, and posited that it might extend over to particles. So if we fire a beam of electrons at a known speed (or equivalently, with a known kinetic energy), then we can compute the wavelength they will exhibit in a diffraction pattern. So for a diffraction grating with slit separations equal to d, we should see "bright fringes" (lots of electron landings) at angles θ_m that satisfy:

$$d\sin\theta = \lambda\,,\ \ 2\lambda\,,\ \ 3\lambda\,,\ldots \qquad \text{where:}\ \ \lambda = \frac{h}{p_e} = \frac{h}{m_e v} = \frac{h}{\sqrt{2mE}} \tag{3.4.3}$$

This wavelength is known as the *de Broglie wavelength*, and indeed we found that this explains the results of the Davisson-Germer experiment.

Alert

We have now seen two differently-defined wavelengths for matter – Compton and de Broglie wavelengths. Don't confuse these with each other! The Compton wavelength is defined purely by the mass of the particle (so it is a fixed number), while the de Broglie wavelength is defined by a particle's momentum (so it changes, depending on how fast the particle is moving).

While both light and matter exhibit these wave & particle properties, for the sake of keeping clear what we are talking about, when we are talking about the wave properties of matter (which we will do a lot from now on), we will avoid confusion by referring to the waves associated with matter as *matter waves*. The reader is cautioned against misinterpreting this moniker, however. "Water waves" are waves for which the medium is water, and "string waves" indicate that the medium is string. It is incorrect, however, to interpret "matter waves" as indicating that the waves are made out of matter. Light travels through no medium whatsoever (there is no luminiferous aether), and the same is true of matter waves.

An Important Feature of Matter Waves

When we discussed double slit interference for the first time, we found that we can only see a significant interference pattern (many bright and dark fringes) if the wavelength of the light is much smaller than the slit separation. This is because the maximum-order of bright fringe is limited by the fact that $\sin\theta \leq 1$:

$$d\sin\theta = m\lambda \quad \Rightarrow \quad m_{max} \leq \frac{d\sin\theta_{max}}{\lambda} = \frac{d}{\lambda} \tag{3.4.4}$$

So if we wish to probe very small dimensions (characterized here by the slit separation), we can't get much information about it if the slit separation is smaller than the wavelength of the light, since we won't be able to make out any features – we can't tell how far apart the slits are if they are so close that no interference pattern is present. This means that we can only make out smaller features (like the separations of atoms in a crystal) with light that has sufficiently-short wavelengths. But very short wavelength light is also very energetic. The wavelength of a particle with mass is significantly shorter than the wavelength of a photon of equal kinetic energy. For example, to look at features on the scale of nanometers (large molecules) requires wavelengths in the same range (about $10^{-9} meters$). To do this with light requires an energy per photon of:

$$E = pc = \frac{hc}{\lambda} = \frac{\left(6.63\times10^{-34}Js\right)\left(3\times10^{8}\frac{m}{s}\right)}{10^{-9}m} = 2.0\times10^{-16}J \tag{3.4.5}$$

An electron of the same wavelength has a kinetic energy of:

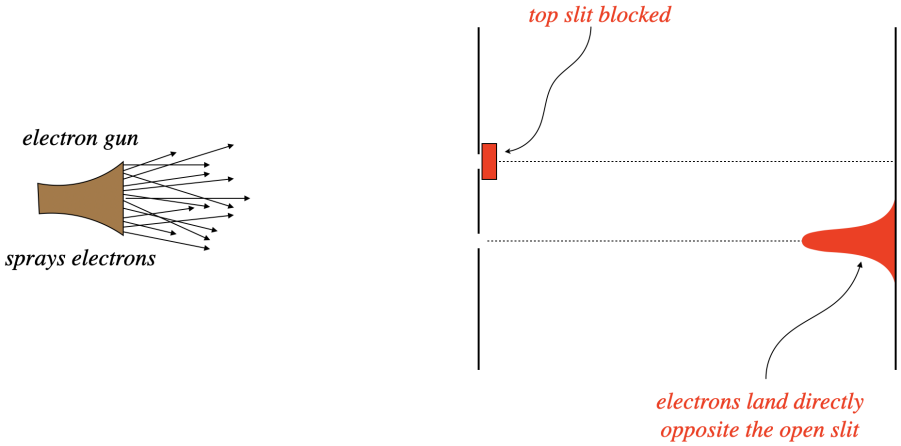
$$E = \frac{p^2}{2m_e} = \frac{h^2}{2\lambda^2 m_e} = \frac{\left(6.63\times10^{-34}Js\right)^2}{2\left(10^{-9}m\right)^2\left(9.11\times10^{-31}kg\right)} = 2.4\times10^{-19}J \tag{3.4.6}$$

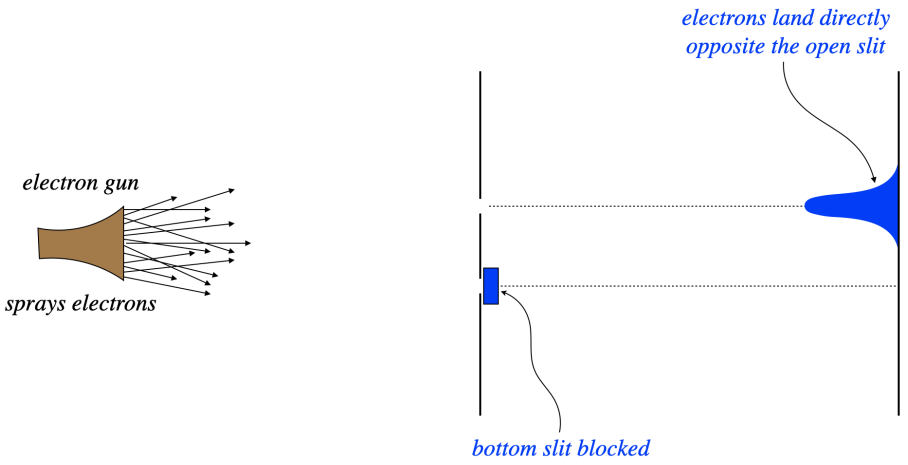
Almost 1000 times as much energy is needed for x-rays than is needed for an electron beam, to achieve the same resolution.

Richard Feynman's Question

What follows is a nice description of our confusing mess as first described by a physicist named Richard Feynman. Consider firing electrons at a double-slit apparatus, with one of the two slits blocked. With one path to the screen, we see a distribution of electron-caused dots on the screen exactly as we would expect – in a cluster centered directly opposite the slit. Next the slits are reversed – the previously-closed slit is opened, and the previous open slit is closed. Unsurprisingly, we see the same result – a cluster of dots on the screen centered across from the slit.

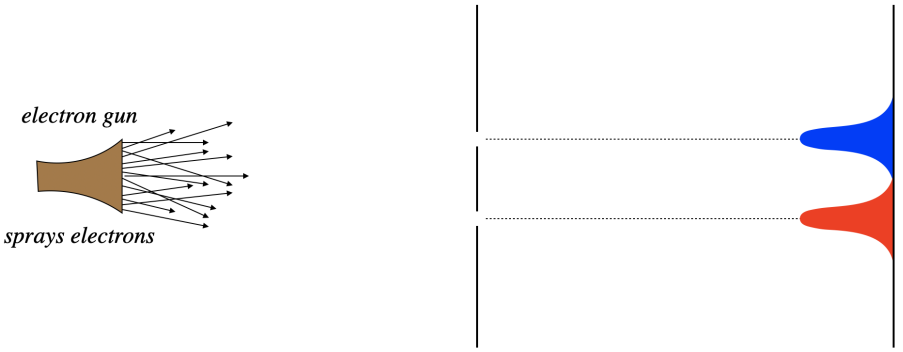
Figure 3.4.1 – Electrons Through a Double Slit with One Slit Blocked





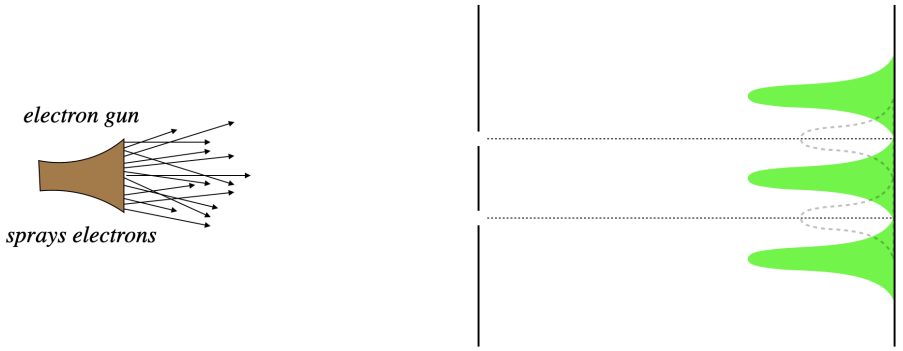
Given these results, the natural prediction of what happens when both slits are open is that we see both of the patterns we saw previously, at the same time. All the electrons that pass through the top slit should end up across from it, and all of those that pass through the bottom slit end up across that that slit.

Figure 3.4.2 – Electrons Through a Double Slit with Both Slits Open (Expected)



But the result (if the slits are spaced appropriately), is completely different. Rather than clusters of dots across from the slits, there are virtually *no dots at all*, and places where we expected very few dots are quite populous. In short, we see an interference pattern!

Figure 6.1.3 – Electrons Through a Double Slit with Both Slits Open (Actual)



The first attempt to explain this is naturally to say that the electrons are interacting with each other after they pass through the slits. But even if the electrons are sent through one at a time, the accumulated landings follow this pattern. So Feynman's puzzling question is, "How does an electron, while going through the bottom slit, know whether or not the top slit is open?" If it is open, it's destination is different than if it is closed, but how can it tell where it is "supposed" to go? Like the case of photons, we can only conclude that the electron somehow passes through both slits at once, in the form of a wave. We will discuss this further in the next section.

3.5: Summarizing this Wave/Particle Mess

Wave-Particle Duality

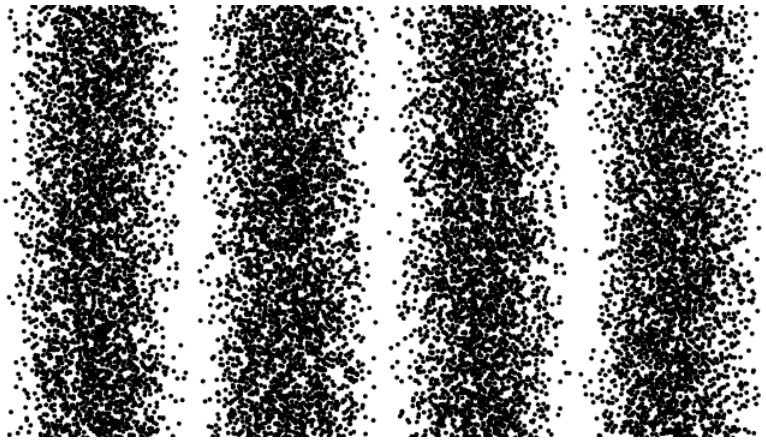
Okay, so we have established that light can behave as a particle (photon) or as a wave (Maxwell). It seems as though how it behaves simply depends upon the context we put it into. If we do typical optics experiments (polaroids, double slit interference, etc.) then light clearly behaves like a wave, but if we do experiments related to the interaction of light with matter (blackbody radiation curve, photoelectric effect, Compton scattering), then it acts very much like a particle. But how can a single entity simultaneously act like two such opposite phenomena? Is light localized into little packets of energy, or is it spread out and able to exhibit interference? IT CAN'T BE BOTH!! This mystery is what physicists refer to as *wave-particle duality*. Giving it a name helps us manage our sanity, as it fools us into thinking it is reasonable. Okay, so we have actually done more than just give it a name, but the reader should understand at the outset that ultimately this is just another one of the incomprehensible wonders that comes from modern physics.

Very Low Intensity Double Slit Interference

It seems like there has to be *some* way to test both the particle and wave nature of light with the same experiment. What most defines the wave nature of light is the phenomenon of interference. This plays no role in any of the experiments we have discussed so far that show the particle nature of light, so let's go back to an interference experiment and see if we can throw-in the particle nature.

The explanation of the photoelectric effect states that light intensity is manifested in the number of photons passing per second. This means that we could, in principle, lower the intensity of the light to the point where only one individual "packet of light" is present. If we fire this photon at a screen where it is detected, it will only make a single dot. If we fire it at a double slit, assuming it makes it through to the other side (i.e. doesn't make a dot on the screen where the slits are cut out), then when it hits the back screen, do we see an interference pattern? The answer is "No, it makes a dot on that screen as well." So that settles it – photons must be particles! Not so fast. We still have to answer where the interference pattern comes from when we should a lot of photons. Maybe they bounce off each other somehow, so that a lot land in some places (bright fringes) and none land in others (dark fringes)? We can test this idea by sending lots of photons, one at a time. Every time a dot appears, we send another single photon. These photons don't land in the same place every time, but they don't land perfectly randomly, either – they form a pattern... a *familiar* pattern. Below is a simulation of what is seen, sped up many times the actual one-dot-at-a-time rate.

Figure 3.5.1 – One-Photon-at-a-Time Through a Double Slit



Our perfectly reasonable idea that light as particles can exhibit wavelike interference behavior by having particles interact with each other fails miserably! The photons travel alone, free from being affected by any other photons, and yet when all the photon landings are aggregated, the interference pattern emerges! We are foiled once again from definitively showing that light is either a particle or a wave.

The only explanation left to us requires a statistical argument: The pattern shown on the screen *must* represent a probability distribution for the landing points of the individual photons. The places dense with dots are places where a single photon has a very

high probability of landing, while those less dense are less probable landing points (and those with no dots have zero probability of a photon landing there!).

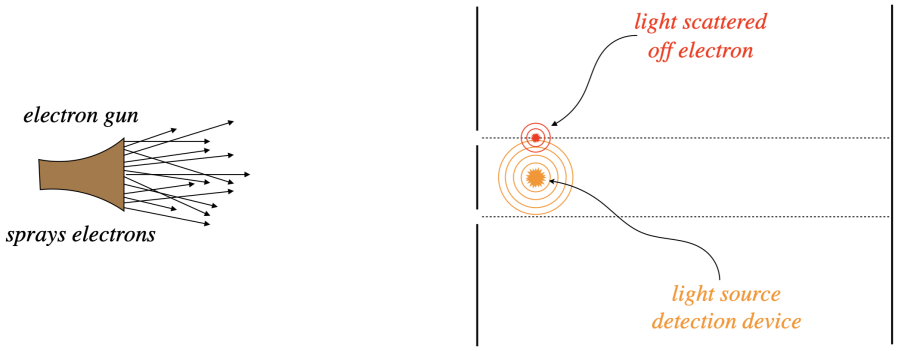
The strange part of all this is that these probabilities appear to obey wave mechanics. That is, when we change the spacing between the slits, or the wavelength of the photons, the pattern changes in a manner that is precisely consistent with wave interference. This puzzle of something as abstractly mathematical as probabilities being subject to the rules that exist for waves is one of the most fundamental aspects of what is known as quantum theory. As we will see, wave-particle duality is not the only place where something we expect to be smooth and continuous (like a light wave's interference pattern) turns out to be discrete (like a single dot on a screen). This probabilistic/statistical interplay between what we see in the big picture as continuous, and what we see in the small scale as discrete leads us to invent new language to describe quantities that doesn't imbue them with inherently "particle" or "wave" properties. We then generically refer to these quantities as "quanta".

We say that until we actually force a quantum to reveal its particle properties (like a dot on a screen), it exists as a mysterious entity that is spread throughout space, and propagates & interferes with itself like a wave. This mysterious wave entity carries with it information about the probabilities of various measurements, most notably the location of the particle.

What if We "Peek"?

In an attempt to solve the puzzle of the double slit from the previous section, we might try to simply watch the electrons as they pass through the slits. To do this, let's put a bright light source between the slits, so that when electrons pass, by the light scatters off them and we see a small flash of light coming from the location of the electron, thereby telling us which slit it went through.

Figure 3.5.2 – Watching the Electrons as They Pass Through the Slits



When we do this, we find we have a problem. The interference pattern disappears, and the previous "expected" pattern of electrons landing either opposite one slit or the other emerges. Apparently we have affected the motion of the electrons after they pass through the slit with our detection device. But of course we have! Light has momentum, and when we scatter it off the electrons, the momenta of the electrons are altered, apparently ruining the interference effect we were trying to study. So the obvious solution? Use light with less momentum, so that it doesn't transfer so much to the electrons. The means we have to use light of long wavelengths.

We previously discussed the limits on resolution of an observation based on the wavelength of light used to make that observation. We noted that wavelengths longer than the slit separation of a double slit will not be able to provide information about the slit separation. Well it turns out that in order to stop affecting the motions of the electrons by using longer-wavelength light, the wavelength must be longer than the space between the slits. So we can use light that doesn't transfer enough momentum to significantly affect the trajectories of the electrons (keeping the interference pattern), but in doing so, the loss of resolution for that longer wavelength makes it impossible for us to determine which slit the electron goes through, which was our whole reason for introducing the light in the first place! Infuriating... and amazing.

Quantum States

One thing this result tells us is that our classical notion of predicting the exact motion of particles using Newton's laws and kinematics must be discarded. Trajectories of particles are inherently probabilistic, and yet there is nevertheless a certain degree of predictability – the interference pattern is quite repeatable. So our task in studying this subject is to determine what physical properties contribute to the observed behavior, and come up with a mathematical model to predict – in a probabilistic manner – the results of experiments with these particles.

We already have an example of how we can use physical properties to predict this probabilistic behavior. The momentum of a particle is directly related to the associated wave's wavelength. So we can make a prediction of where no particles will land on the screen (dark fringes) by knowing how fast the particles were moving (and of course their masses) when we shot them at the double slit.

Given the weirdness of these quanta, it's not clear what all the properties are that define the probabilities we seek to predict. When we studied thermodynamics in Physics 9HB, we defined something we called a "thermodynamic state". This was the equilibrium condition of a system, which was completely defined by several variables, like temperature, pressure, and volume. We will now define what we call a *quantum state*. Unlike a thermodynamic state, where knowing enough thermodynamic quantities to define the state tells us the values of other quantities (e.g. knowing the number of moles, volume, and pressure of an idea gas tells us exactly its temperature), in quantum physics, we can usually only know *probabilities* of the state's values being measured.

If we are interested in the position of a particle, we sort the quantum state into a collection of probabilities for every point in space. We will simplify this discussion by restricting ourselves to a "space" on one dimension. The information that the quantum state holds about the probabilities of various positions is called the *wave function*. In the language of the bras and kets discussed in Section 1.6, the wave function is the "component" of the quantum state vector that multiplies the unit vector defining position x:

$$\psi(x) = \langle x|\psi\rangle \tag{3.5.1}$$

We have greatly simplified things here (for example, we have not made any mention of how the quantum state evolves through time), but the general idea is this: The wave function $\psi(x)$...

- ... carries information about the probability of the particle being measured ("making a dot") at position x,
- ... obeys the superposition principle, which means it can interfere with itself to create things like double-slit patterns and standing waves,
- ... has wave properties like wavelength (or a combination of wavelengths, if it is a superposition of waves) that correspond to physical properties like momentum,
- ... depends upon the physical situation that the particle finds itself in (like being acted-upon by external forces)

The topic of *quantum mechanics* is the study of solving for this wave function in various situations, and using it to make probabilistic predictions of what will be observed.

CHAPTER OVERVIEW

4: The Universe is Inherently Probabilistic

4.1: Basics of Probability Theory

Probability Distributions

Whenever we consider an outcome to a random event, the probability of that outcome is the ratio of its "measure" to the measure of all possible outcomes. Frequently this measure is computed purely by counting, and is particularly simple if each outcome is equally likely. Sometimes the counting is a bit more complicated, because the desired probability spans a group of distinguishable results. For example, if one asks for the probability of throwing a 7 on a standard pair of 6-sided dice, this "outcome" can occur in many ways: The first die can result in 1 and the second in 6, the first die 6 and the second 1, the first 5 and the second 2, and so on. There are 6 distinct rolls that result in this same outcome, giving it a measure of 6. The total number of possible rolls is the product of the number of possible results on the first die and the number of possible results on the second die, or 36. The ratio of the measure of the desired outcome to the measure of all outcomes is 6/36 = 1/6.

If the outcome of a roll of two dice is defined as the total on both dice, then all outcomes are not equally-probable. For example, there are more ways to roll a total of 6 (five ways) than a total of 5 (four ways). The map of probabilities for the various possible outcomes is called the *probability distribution*. For two dice, there are eleven possible outcomes, and the probability distribution for these outcomes are shown below.

Figure 4.1.1 – Probability Distribution for Sum of Two Six-Sided Dice

This probability distribution involves different probabilities for different outcomes. Those (like for the roll of a single die) that provide the same probability for all outcomes are called *uniform*. Quite often when one knows of no mechanism that would cause results to clump into certain outcomes, the first guess at the probability distribution is uniform. This assessment can change either with increased knowledge of the randomizing mechanism (e.g. a single die is found to be weighted unsymmetrically), or data indicating that the original assumption was likely bad (e.g. a single die comes up one number much more often than random chance would indicate).

Mutually-Exclusive Outcomes

In the case of the roll of a pair of dice, clearly it is impossible to get both a 6 and an 8 on the same roll. The roll of a 8 excludes the possibility of the roll of an 8, and a roll of an 8 excludes the possibility of the roll of a 6. Two outcomes of these kinds are referred to as *mutually-exclusive*.

If we enumerate all of the mutually-exclusive outcomes for a random event, the one and only one of them must occur. The possible rolls of two dice are the numbers 2 through 12. If we add up all of the probabilities for these outcomes, we get a sum equal to 1. We can ask what the probability is that the random event will result in either outcome A *or* outcome B. If these outcomes are mutually-exclusive, then the probability of one or the other occurring is the sum of the probabilities:

$$P(A \text{ or } B) = P(A) + P(B) \qquad \text{A and B mutually exclusive} \tag{4.1.1}$$

One must be very careful to only use this sum of probabilities under these restrictions. Not *all* probabilities that involve "or" is a simple sum. For example, one could ask what the probability is that for a roll of two dice, the red die comes up 2, or the green die comes up 2. The probability of a single die coming up 2 is $\frac{1}{6}$, so one might guess that the probability of one or the other coming up 2 is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, but it is not! The roll of a 2 on the red die does not exclude the roll of a 2 on the green die. As these are not mutually-exclusive, you cannot add their probabilities.

Independent Outcomes

The opposite case of mutually-exclusive outcomes are *independent outcomes*. As suggested by the name, two outcomes that are independent have no effect over one another. The case above of the 2 coming up on the red die and the 2 coming up on the green die is an example of two independent outcomes. As we saw above, we cannot sum probabilities for "or's" in these cases, but there is a parallel bit of math we can do.

When the dice are rolled, we can use the individual probabilities to compute the probability of *both* independent events occurring. This is a simple matter of multiplying the probabilities:

$$P(A \text{ and } B) = P(A) \cdot P(B) \qquad \text{A and B independent} \tag{4.1.2}$$

In the case of rolling 2's on both dice, the probability is the product of the two probabilities: $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

Note that this also makes it possible to *properly* compute the probability of a 2 on either the green or the red (or both). We do this using the "and" property as follows: The probability that one die misses a 2 is $\frac{5}{6}$, so the probability that both dice miss 2's is $\frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36}$. But if both dice missing 2's doesn't occur, then that means at least one of the dice is a 2. Since these two cases represent all the outcomes, the sum of their probabilities is 1, which means that the probability of the red die rolling a 2 or a green dies rolling a 2 (or both rolling 2's) is $1 - \frac{25}{36} = \frac{11}{36}$. Note that this is not quite equal to the $\frac{1}{3}$ result found in the erroneous calculation earlier.

Expectation Values

Where probabilities become particularly important is when the outcomes are accompanied by measurable results. Continuing with our pair of dice example, let's suppose that two people wager on the outcome of a roll. If the same wager is made over and over many times, the outcomes can be summed together and divided by the number of rolls to get an average outcome per roll. This average of a single outcome need not be one of the possible outcomes of a single roll, but is referred to as the *expectation value* of a single event.

The way to use probabilities to compute expectation values is as follows: Multiply the probability of every outcome by the value of that outcome, and sum all of these products:

$$\langle \omega \rangle = P(A) \cdot \omega(A) + P(B) \cdot \omega(B) + \ldots \tag{4.1.3}$$

Let's look at an example of an expectation value for our dice roll model...

Ann and Bob agree that Ann will receive \$4 from Bob if the result of a two-dice roll is a 4 or a 7, and Bob will receive \$3 from Ann if the result is a 6 or an 8. Neither wins any money if anything else is rolled. Let's compute the expectation value of what Ann will receive per roll. The results 4 and 7 are mutually-exclusive, so we can add their probabilities to get the probability of Bob winning: $\frac{3}{36} + \frac{6}{36} = \frac{1}{4}$. The probability of Bob losing is similarly: $\frac{5}{36} + \frac{5}{36} = \frac{5}{18}$. The remainder is the probability that no one wins: $1 - \frac{9}{36} - \frac{10}{36} = \frac{17}{36}$. Now multiply these probabilities by the outcomes for Ann to get her expectation value:

$$\langle \omega \rangle = \left(\frac{1}{4}\right)(+\$4) + \left(\frac{5}{18}\right)(-\$3) + \left(\frac{17}{36}\right)(\$0) = +\frac{1}{6}(\$1) \tag{4.1.4}$$

4.2: Continuous Probability Distributions and Probability Density

Infinite Number of Outcomes

There is no reason why all probability distributions must be discrete as it is for two dice. Probability distributions on a continuum are also possible. The probability of a blindfolded dart thrower hitting various positions on a dart board could be an example of a two-dimensional continuous probability distribution.

The key feature of probability on a continuum is that one can no longer say that a given outcome has a specific probability. If one selects a number at random from 0 to 1, the probability of hitting exactly a predicted number is zero, as there are uncountably-many choices. Even though one of those numbers is selected, its probability of being correctly-guessed was zero.

In such cases where the outcomes lie on a continuum, we need a different way to express probabilities – we need to express them in a *range*. So rather than talk about the probability of an outcome being exactly equal to x, we define a probability of lying between x_1 and x_2. If the probability density on the continuum is uniform, then the calculation of the probability of lying within a range is easy. If, for example, all the outcomes of a random number lie on a number line between 0 and 8, then the probability of a single outcome occurring between 1.2 and 3.6 is the ratio of the size of the target range and the size of the full range: $P\left(1.2 \leftrightarrow 3.6\right) = \frac{3.6-1.2}{8} = 0.3$.

But what if the probability distribution is not uniform, that is, what if the outcomes at some places in the continuum are more probable than others?

Probability Density

If we have a continuous probability distribution (of any dimension), then the measure for any individual result is actually zero, as there are infinitely-many possible outcomes. However, this doesn't make all the outcomes equally likely, because they may have different relative measures. For example, if the probability of one outcome is P and the probability of a second outcome is $3P$, then the ratio of these outcomes shows that the latter outcome is three times more likely than the former, even in the limit as P goes to zero. Also, the sum of the infinite number of zero-probability outcomes still must equal one. We assure that this works properly by representing the continuous probability distribution with a *probability density function*.

As with any other density function we have encountered (such as mass density), the idea is to measure the relative weightings at various positions. For a line of mass along the x-axis with a mass density of $\lambda\left(x\right)$, the infinitesimal amount of mass found in the tiny slice between positions x and $x+dx$ is given by $dm = \lambda\left(x\right)dx$.

Figure 4.2.1 Amount of Mass In an Infinitesimal Section in Terms of Density

Now imagine that instead of a line of matter with varying mass density, we were talking about a particle bouncing back-and-forth within an opaque tube. The particle could be anywhere within the tube, and its probability of being between x and $x+dx$ is infinitesimally small. But we can describe the probability of it being in that region in terms of the probability density function $\mathcal{P}\left(x\right)$ in the same way as we did for mass:

$$dP\left(x \leftrightarrow x+dx\right) = \mathcal{P}\left(x\right)dx \tag{4.2.1}$$

Then the probability of it lying within a finite range is just the sum (the outcomes are mutually-exclusive) of all of these infinitesimal probabilities:

$$P\left(x_1 \leftrightarrow x_2\right) = \int_{x_1}^{x_2} dP = \int_{x_1}^{x_2} \mathcal{P}\left(x\right)dx \tag{4.2.2}$$

Normalization

A universal truth of probability theory is that when the result of a random event occurs, it must land within the universe of possible outcomes. Mathematically, this means that the sum of the probabilities of all possible outcomes must be 1. This can be confirmed for the case of the roll of two 6-side dice by summing all of the probabilities in Figure 4.1.1.

What distinguishes the various probabilities from each other are their *relative* measures. In the example of the two dice, the probability of throwing a 7 is twice as great as throwing a 4 or a 10. We can determine these measures by comparing the number of ways the results can occur (six ways for the 7 versus three ways for the 4 and 10), but if we want to be able to properly use the probability distribution, we must divide all these measures by the sum of all measures so that the new sum is 1. This process is called *normalization*.

Imposing the normalization condition on a probability density function requires that:

$$1 = \int_{\text{all } x} \mathcal{P}(x)\, dx \tag{4.2.3}$$

In the work that follows, 'x' will usually (but not always!) refer to an actual position in a one-dimensional space, so "integrating over all x" means that the normalization condition is typically:

$$1 = \int_{-\infty}^{+\infty} \mathcal{P}(x)\, dx \tag{4.2.4}$$

Expectation Value

To complete our extension of the previous section to the case of a continuum of outcomes, we have to address expectation values. If there are infinitude of possible outcomes because they are distributed on a continuum, then the sum given in Equation 4.1.3 is a sum of the product of the infinitesimal outcome probabilities multiplied by the values for each of the outcomes:

$$\langle \omega \rangle = \int_{-\infty}^{+\infty} \mathcal{P}(x)\, \omega(x)\, dx \tag{4.2.5}$$

It is important to note that the expectation value is, in statistics terms, the *mean* of the distribution (as opposed to the mode and median, two other statistical measures of the "center" of a distribution), which means that like the discrete case, this value is not necessarily one of the possible outcomes.

Example 4.2.1

A block vibrates on a frictionless horizontal surface while attached to a spring with spring constant k. The maximum distance that the mass gets from the equilibrium point is x_o. A radar gun measures the speed of the block at many random times, and these speeds are combined with the mass of the block to compute the block's kinetic energy. Find the average kinetic energy measured.

Solution

There are several ways to approach this. We will take the brute-force method here, to emphasize the mathematical details of the probability density integral. We start by determining the probability of the block being between x and $x + dx$ at any random moment (with x measured from the equilibrium point of the spring). First, it should be clear that the probability density is not uniform – the block spends longer near the extreme ends of the oscillation than near the center, because it is moving slower near the endpoints. The probability of being in the tiny range dx will be the ratio of the time it spends there (which we'll call dt) to the time it spends going from one end of the oscillation to the other (half a period, $\frac{1}{2}T$):

$$P(x)\, dx = \frac{dt}{\frac{1}{2}T} \quad \Rightarrow \quad P(x) = \left(\frac{2}{T}\right)\frac{dt}{dx} = \frac{2}{vT}$$

Plugging this into the expectation value equation for kinetic energy gives:

$$\langle KE \rangle = \int_{-x_o}^{+x_o} P(x) \left[\frac{1}{2} m v^2 \right] dx = \frac{m}{T} \int_{-x_o}^{+x_o} v \, dx$$

Clearly the velocity of the block changes with respect to x, so v cannot be pulled out of the integral. The function of x that we plug in to v is found by noting that the total energy of the system remains constant, and equals the potential energy at the extreme points of the oscillation:

$$E = \frac{1}{2} m v^2 + \frac{1}{2} k x^2 = \frac{1}{2} k x_o^2 \quad \Rightarrow \quad v(x) = x_o \sqrt{\frac{k}{m}} \sqrt{1 - \left(\frac{x}{x_o} \right)^2}$$

Plugging this into the integral and making the substitution $u \equiv \frac{x}{x_o}$ gives:

$$\langle KE \rangle = \frac{m x_o^2}{T} \sqrt{\frac{k}{m}} \int_{-1}^{+1} \sqrt{1 - u^2} \, du$$

The reader that wants to do every step of the math can perform the integral with a trig substitution, but looking it up is also fine – it comes out to equal $\frac{\pi}{2}$. All that remains is to use the period of oscillation for this simple harmonic oscillator in terms of the mass and spring constant:

$$T = 2\pi \sqrt{\frac{m}{k}} \quad \Rightarrow \quad \langle KE \rangle = \frac{1}{4} k x_o^2$$

Note that the average kinetic energy is half the total energy, which means the average potential energy is the same – on average the energy is split evenly between the two modes.

4.3: The Uncertainty of Random Outcomes

Quantifying What We Don't Know

When dealing with outcomes of random events, we can find expectation values, which accurately predict the average result over a large number of repeated attempts, but when it comes to a single attempt, this number is nothing more than an educated guess. So while we are confident about the average over a large (infinite) number of trials, we don't have such confidence about a single trial, and it would be good to know just how close we expect our guess to be to the actual result. We can quantify this level of *uncertainty* mathematically.

The expectation value gives us an average result for many trials, but we can also have a look at how much the results spread themselves out. If the results spread very wide, then selecting from the wide range of results could land very far from the mean, which indicates that we are quite uncertain of our estimate. If the results spread in a narrow range, then any given result is likely to be quite close to our estimate, and our uncertainty in our guess is low. We can compute a value for this uncertainty by checking how far every possible outcome is from our expectation value guess, and adding these "deviations" together. This measure is called *standard deviation*.

The computation of standard deviation goes like as follows. We will start here with the case where there are discrete results, and then generalize to the case where a probability density is needed. First, we need to know how far each possible individual result, ω_i, is from the mean of all the results, $\langle\omega\rangle$:

$$\text{separation of the } i^{th} \text{ result from the mean} = \omega_i - \langle\omega\rangle \tag{4.3.1}$$

We would like to know the "average separation" over all the results, but how do we define such an average? If we just take an average of the differences given above, it would come out to equal zero. The proof of this is easy:

$$\langle\omega_i - \langle\omega\rangle\rangle = \frac{1}{N}\sum_{i=1}^{N}\left[\omega_i - \langle\omega\rangle\right] = \frac{1}{N}\sum_{i=1}^{N}\omega_i - \frac{\langle\omega\rangle}{N}\sum_{i=1}^{N}(1) = \langle\omega\rangle - \langle\omega\rangle = 0 \tag{4.3.2}$$

The problem is that these separations are both positive and negative, but to measure the spread, we don't care in which direction the deviation from the mean is. We could define the average deviation of the results from the mean as the average of the absolute values of the separations, but for rather mathematically complex reasons, it turns out that this is not the best definition. We won't go into details here, except to say that it is more useful to give a higher weighting to deviations as they get farther from the mean (the absolute value method weights all deviations equally).

The "standard" deviation that we calculate also removes the problem of negative deviations, but also weights separation from the mean more as it becomes greater. It does this by *squaring* the separation of every result from the mean, averaging those squares, and then taking the square root of the sum. In other contexts where the mean is clearly zero (such as the current in an AC circuit), this is a measurement of the average magnitude of the value, and is often referred to as the *root-mean-square*, or *rms* value, for reasons that are obvious now that we know how it is calculated.

Let's summarize the calculation of standard deviation before writing out the formula.

- Start with the full set of outcomes, ω_i, and their accompanying probabilities, P_i.
- Calculate the mean (expectation value) of the outcomes, using Equation 4.1.3.
- Calculate the separation of every outcome from the mean, using Equation 4.3.1.
- Square all of these separations.
- Find the mean of all these squares (add them all together and divide by the total number).
- Take the square-root of this mean.

$$\Delta\omega = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\omega_i - \langle\omega\rangle\right]^2} \tag{4.3.3}$$

This formula can actually be cast into another extremely useful form – so useful that we will end up using this alternative form pretty much exclusively. To get to it requires only a little bit of algebra: Expand the square inside the sum, and use the facts that $\langle\omega\rangle$ is a constant value, not dependent on i, and $\frac{1}{N}\sum_{i=1}^{N}\omega = \langle\omega\rangle$. The result is:

$$\Delta\omega = \sqrt{\langle\omega^2\rangle - \langle\omega\rangle^2} \tag{4.3.4}$$

The description of this process is easy to put into words: "Compute the average of the squares of the outcomes, subtract the square of the average outcome, and take the square root." We will see that this form is especially useful when we go to the case of a continuum of possible outcomes, which we do now...

Uncertainty for a Continuum of Outcomes

We already know how to compute a mean using a probability density (Equation 4.2.5). All we have to do to calculate the uncertainty is compute two of these expectation value integrals (one for the value itself, and one for the square of the value) and then plug the results into Equation 4.3.4.

$$\left.\begin{aligned}\langle\omega\rangle &= \int_{-\infty}^{+\infty} \mathcal{P}(x)\,\omega(x)\,dx \\ \langle\omega^2\rangle &= \int_{-\infty}^{+\infty} \mathcal{P}(x)\,[\omega(x)]^2\,dx\end{aligned}\right\} \quad\Rightarrow\quad \Delta\omega = \sqrt{\langle\omega^2\rangle - \langle\omega\rangle^2} \tag{4.3.5}$$

Example 4.3.2

In Example 4.2.1, symmetry demands that the average position of the block is the origin. Find the uncertainty in the block's position.

Solution

With an average position of $\langle x\rangle = 0$, Equation 4.3.4 tells us that the uncertainty in the position of the block is:

$$\Delta x = \sqrt{\langle x^2\rangle}$$

Now we can plug into the integral using the density function we found in Example 4.2.1, but that is reinventing the wheel. It's simpler to use what we found in that example:

$$\langle PE\rangle = \frac{1}{4}kx_o^2 \quad\Rightarrow\quad \left\langle \frac{1}{2}kx^2 \right\rangle = \frac{1}{4}kx_o^2 \quad\Rightarrow\quad \langle x^2\rangle = \frac{1}{2}x_o^2$$

This gives us the uncertainty of x:

$$\Delta x = \frac{1}{\sqrt{2}}x_o$$

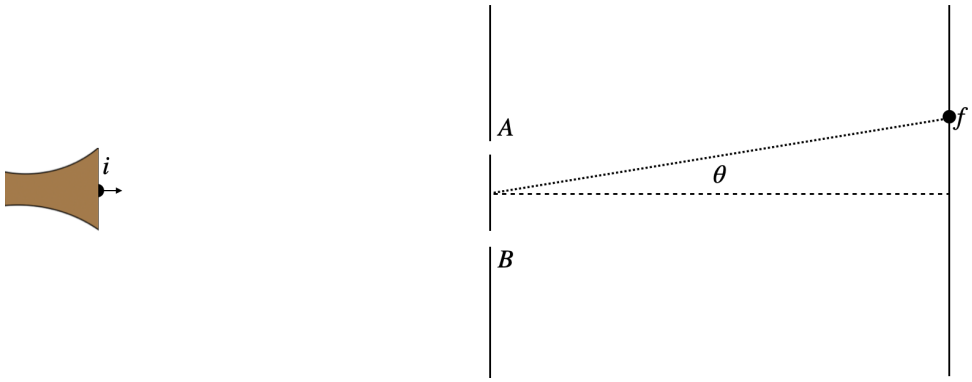
4.4: Physical Measurements with Random Outcomes

Probability Amplitude

We have discussed how light and matter both behave light particles and waves, and the fact that combining the wave nature with the observance individual dots on a screen leads us to the inescapable truth that nature behaves probabilistically on a fundamental level. We said this was because the interference pattern occurred even when we sent one particle at a time (i.e. we waited for the dot to appear before sending another particle), which means that regions with higher densities of dots must be more probable landing points than those with low densities of dots. Let's apply some of the tools of probability theory to the double-slit experiment in order to construct a mathematical model for what is happening.

We'll start by labeling a few parts of the double-slit experiment. We'll label a starting point of the electron or photon (the source of the beam) with an 'i', and the eventual landing point (dot on the screen) with an 'f'. The two possible paths (slits) we will label with an 'A' and a 'B'.

Figure 4.4.1 – Defining Paths for a Particle Through a Double-Slit



Returning to the idea that this particle has a localized nature (it makes dots on the screen), it seems reasonable to conclude that this dot *either* passed through slit A *or* through slit B, and that these two choices are mutually exclusive. From what we know of probability theory (Equation 4.1.1), we can write the probability of getting from i to f as the sum of the probabilities of making this journey through slit A and through slit B:

$$P\left(i \to f\right) = P\left(i \to A \to f\right) + P\left(i \to B \to f\right) = P_{\text{through A}}\left(\theta\right) + P_{\text{through B}}\left(\theta\right) \tag{4.4.1}$$

But if we close slit B, the probability $P\left(i \to A \to f\right)$ doesn't change. The only difference is that the particles that were previously going through slit B get blocked by the barrier we put there. So using this math leads to the incorrect prediction depicted in Figure 3.4.2.

The problem is with destructive interference. Probabilities (and probability densities) are always positive, so they cannot cancel each other, as these seem to do. So to explain this experiment, we must marry the purely-positive nature of probabilities with the potential for destructive interference. Well it turns out that we've had the idea for this all along! We found long ago two things:

1. The intensity pattern is directly related to the probability – the more densely-packed the dots are in a region of the screen, the higher the probability must be that a single particle will land in the region.
2. Intensity of a wave is proportional to the *square of its amplitude*.

We therefore infer the existence of a *probability amplitude*. This is the amplitude of the wave (whether it is an EM wave or a matter wave), that can interfere with itself in a double-slit apparatus. The phase of the wave can turn this amplitude into a positive or negative (or actually, as we will see later, a complex) quantity, which allows it to result in destructive interference. The square of this value (which becomes a bit more complicated when we get to complex-valued amplitudes) is the probability density, which remains strictly positive, as it must.

Because of the principle of superposition of waves, these probability amplitudes (adjusted by the phases of their associated waves) are the quantities that we add for the two paths, rather than the probabilities. The paths are *not* mutually exclusive – the wave passes through both slits, not strictly one or the other. If we call the wave that arrives at f (deflects through θ) from i through slit A '$\psi_A\left(\theta\right)$' and the wave that goes through slit B '$\psi_B\left(\theta\right)$', then the total wave is:

$$\psi_{tot}(\theta) = \psi_A(\theta) + \psi_B(\theta) \tag{4.4.2}$$

The probability density at that position on the screen is then the square of this total (in anticipation of these probability amplitudes being complex, we'll use the proper description of these squares here, but don't worry about this notation yet):

$$\mathcal{P}(\theta) = |\psi_{tot}(\theta)|^2 = |\psi_A(\theta) + \psi_B(\theta)|^2 = |\psi_A(\theta)|^2 + |\psi_B(\theta)|^2 + 2\psi_A^*(\theta)\psi_B(\theta) \tag{4.4.3}$$

The last term here is where there is potential for negative numbers. If not for this "overlap" term, then the mutually-exclusive assumption would be correct.

Summarizing the Path from Physical Properties to Probabilistic Predictions

We have now seen three different quantities that include the word "probability" in their names, and it is useful to have a look back to make sure we have them straight, and clarify how they fit into the physics. We will do this by outlining the steps of how we "do" quantum physics...

- Physical features of the system are determined. These are things like the separation of two slits, the gap sizes of each slit, directions of polarization of polaroids, electrical forces on charged particles, etc.
- The physical features have an effect on the wave function and its boundary conditions.
- Use superposition of wave functions to obtain a single *probability amplitude*:

$$\psi_{tot}(x) = \psi_1(x) + \psi_2(x) + \ldots \tag{4.4.4}$$

- Square the probability amplitude to get the *probability density*:

$$\mathcal{P}(x) = |\psi_{tot}(x)|^2 \tag{4.4.5}$$

- Use the probability density over a range of values to obtain the *probability* that a value lands in a range:

$$P(x \leftrightarrow dx) = \mathcal{P}(x)\,dx = |\psi_{tot}(x)|^2 dx \tag{4.4.6}$$

- Use the probability density to compute expectation values and uncertainties of physically-measurable quantities:

$$\langle \omega \rangle = \int_{-\infty}^{+\infty} \omega(x)\,|\psi_{tot}(x)|^2 dx \tag{4.4.7}$$

Of course there are many details left out here, the most notable being the grief and heartache that goes into finding the wave function from the physical conditions. But there are also mathematical details to keep in mind, such as making sure that the probability density is normalized and a few other things we will discuss soon. But this outlines the main procedure we will follow.

Observables Other than Position

So far we have focused only on the probabilistic nature of measurement of position, but this is certainly not the only observable for which quantum mechanics provides randomness. We have so far only dealt with matter and light with a single wavelength (in one dimension, these are plane waves). Such quanta will not result in a probabilistic measure of momentum or kinetic energy – a single wavelength means a single value for these quantities. But as we noted all the way back in Equation 1.1.16 (and used over and over since then, most notably in Fourier analysis), a solution to the wave equation can be linear combination of many waves, which can all have different wavelengths. This means that the wave function of a single quanta can actually be a linear combination of many single-wavelength waves. If the particle is confined, then its wave is a standing wave, and is a sum of single wavelength harmonics (i.e. is a Fourier series). But even if it is free from confinement, its wave can be a mix of waves of many wavelengths.

What would the momentum of such a particle be, if its de Broglie wavelength is not uniquely-defined? That's exactly the point – it has no specific momentum! We can of course make a measurement of such a particle's momentum, and we will get a value that is associated with one of the many wavelengths that make up this wave. The wavelength that gets chosen from the collection is selected at random, and each wavelength has its own probability of being selected. Momentum (and with it, kinetic energy) is – like position – determined probabilistically!

Dirac Brackets Again

We return now to those enigmatic bras and kets that first appeared in Section 1.6 and later made a brief appearance in Section 3.5. We said that the quantity $|\psi\rangle$ is an abstract vector that contains all the information of the quantum state, and that we can extract this information from it by taking "dot products". Now we have the language to put this together. Suppose we wish to know the probability that the particle in the quantum state $|\psi\rangle$ will be found to be at position x. The quantum state of being precisely at x (i.e. seeing the dot on a screen at x), we define as $\langle x|$. The full quantum state of the particle includes the possibility of the particle being anywhere, but the dot product of this full state with the precise state of location at x is the probability amplitude of the particle being found there:

$$\text{probability amplitude of measuring particle's position to be } x = \langle x|\psi\rangle \tag{4.4.8}$$

Given that there is a separate value of this probability amplitude at every position, this can be written as a function (a wave function), as we already stated in Equation 3.5.1:

$$\langle x|\psi\rangle = \psi(x) \tag{4.4.9}$$

But this doesn't only apply to position! The quantum state vector also contains information about quantities like momentum. The quantum state of a particle having a precise momentum (which we know manifests as a single-wavelength wave) we will call $\langle p|\psi|$. Then the probability amplitude of measuring the particle's momentum to be p is:

$$\text{probability amplitude of measuring particle's momentum to be } p = \langle p|\psi\rangle = \phi(p) \tag{4.4.10}$$

[Note: While not strictly necessary since the context is usually clear, it is traditional to use a different symbol for a wave function expressed with momenta than is used for positions. So we will typically use $\psi(x)$ and $\phi(p)$.]

All of the same machinery we developed for calculating probabilities for positions applies equally to wave functions written in terms of momentum. That is, the probability of measuring a particle's momentum to be between p_1 and p_2 is:

$$P(p_1 < p < p_2) = \int_{p_1}^{p_2} |\phi(p)|^2 dp \tag{4.4.11}$$

And the expectation value of momentum is:

$$\langle p\rangle = \int_{-\infty}^{+\infty} p\,|\phi(p)|^2 dp \tag{4.4.12}$$

We get a nice bonus in the case of momentum, in that we also get kinetic energy, since $KE = \frac{p^2}{2m}$:

$$\langle KE\rangle = \int_{-\infty}^{+\infty} KE\,|\phi(p)|^2 dp = \frac{1}{2m}\int_{-\infty}^{+\infty} p^2\,|\phi(p)|^2 dp = \frac{1}{2m}\langle p^2\rangle \tag{4.4.13}$$

What About Time?

The reader may be wondering about what happened to the time element for waves – aren't wave functions supposed to look like "$f(x,t)$"? Yes, of course! We got away from worrying about the time portion of the wave function because we were discussing static interference patterns. Although the wave function that results in a double-slit pattern evolves with time, the result itself comes out to be time-independent, so we were able to ignore the effect of the time contribution. As it turns out, we will be able to do this quite a lot in the chapters to come, largely because of the separation of variables trick we first discussed in Section 1.2.

But there is another aspect of this that should not be overlooked. An interference pattern is often *not* static. For example, a standing wave on a string is certainly not static – the string vibrates with time! But when it comes to probability amplitude, it is exactly that – the *amplitude* – that contributes to the critical probability density. Aside from the nodes, every point on a string with a standing wave harmonic is moving, but all of these points have amplitudes that are constants in time. The *position* of that piece of string is changing, but its *amplitude* (its maximum displacement) remains constant when the standing wave is a harmonic. It is this amplitude that comes into play in quantum probabilities, so even though the wave function may be changing with time, the

probabilities associated with different positions may remain fixed. We will later see what physical properties must exist for this to be true.

4.5: Incompatible Measurements

Plane Waves

We have spent a lot of time talking about particles associated plane waves, because they are easy to conceive – they have a single, specific wavelength, and therefore a definite momentum. These are particles free from any forces, moving at a constant, well-defined speed. Easy, right? Well, let's take a look at this wave function in terms of locating the position of the particle. Choosing a cosine function to describe this wave function moving in the $+x$-direction, we have:

$$\psi(x,t) = A\cos\left(\frac{2\pi}{\lambda}x - 2\pi f t\right) \tag{4.5.1}$$

Let's simplify this discussion by looking at the wave only at time $t = 0$:

$$\psi(x) = A\cos\left(\frac{2\pi}{\lambda}x\right) \tag{4.5.2}$$

If we form a probability density from this probability amplitude, we get:

$$\mathcal{P}(x) = |\psi(x)|^2 = A^2\cos^2\left(\frac{2\pi}{\lambda}x\right) \tag{4.5.3}$$

Okay, let's normalize our probability density (i.e. find the value of A):

$$1 = \int_{\text{all } x} \mathcal{P}(x)\,dx = \int_{-\infty}^{+\infty} A^2\cos^2\left(\frac{2\pi}{\lambda}x\right)dx \tag{4.5.4}$$

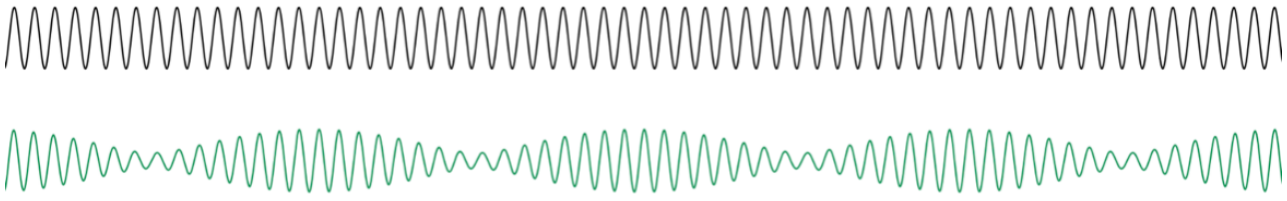
Uh-oh. We have a problem. This integral blows up, making $A = 0$, and $\mathcal{P} \equiv 0$. This makes no sense, what is going on here?

Actually, it does make sense – with a wave function that has the same amplitude along the entire x-axis, the particle must be equally-probable to be found anywhere, so if we look for it in any finite interval, the measure of that interval is zero compared with the measure of the remaining infinite space where it can be found. A simple way to express this is that *we have no idea where the particle is*. This complete lack of knowledge about the particle's position was the small price we had to pay for knowing the particle's momentum precisely. We will see here that this price cuts both ways.

Localizing Free Particles

We have a hint that the precise knowledge of a particle's momentum goes together with a complete lack of knowledge of its location, so let's see if reducing what we know about momentum (or equivalently, wavelength) has the effect of improving our ability to discern the particle's position. We'll start simple: Let's see what happens when we superpose two plane waves with different wavelengths.

Figure 4.5.1 – Superposition of Two Plane Waves



The top graph depicts a plane wave. The one below it is what happens when another plane wave with a slightly shorter wavelength (the wave number $k = \frac{2\pi}{\lambda}$ is larger) is superposed with it. The contribution of this second wave to the superposition is slightly less as well (i.e. its amplitude is smaller than that of the original wave):

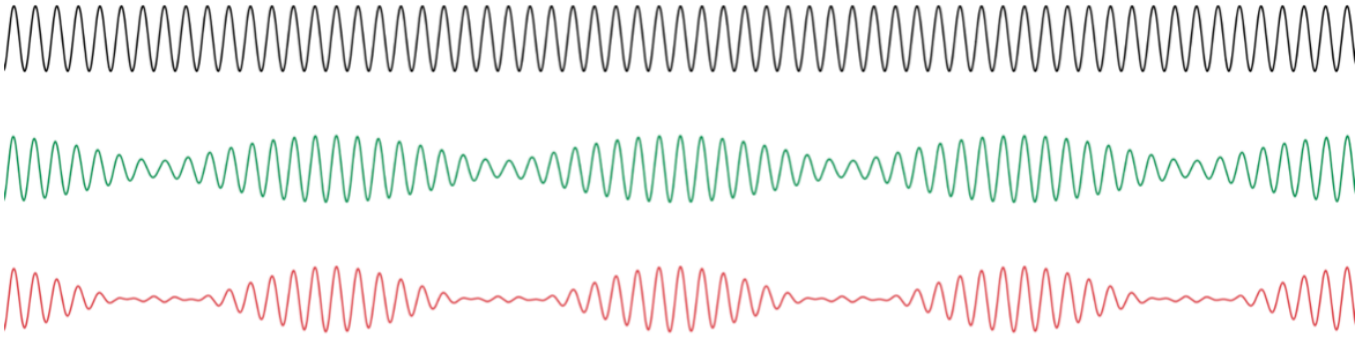
$$\psi(x) = A\cos kx$$

$$\psi(x) = A\cos kx + 0.6A\cos(k + 4\delta k)x$$

The first thing that jumps out is the way that the amplitude (not the displacement!) varies when you evaluate it at different places on the x-axis. In probability terms, this means that the probability of finding the particle in a region near the maximum bulges ("antinodes") is quite high compared to finding it near the narrower regions ("nodes"). While this wave is still infinitely-long, and our knowledge of the position of the particle is still zero, in a *relative* sense, we have a better sense of where the particle is than when we were dealing with a single harmonic wave function.

If we can do this much just by using two wavelengths, perhaps we can improve things even more by adding some additional harmonic waves. If we choose a third plane wave appropriately, we find that the bulges become more defined:

Figure 4.5.2 – Superposition of Three Plane Waves

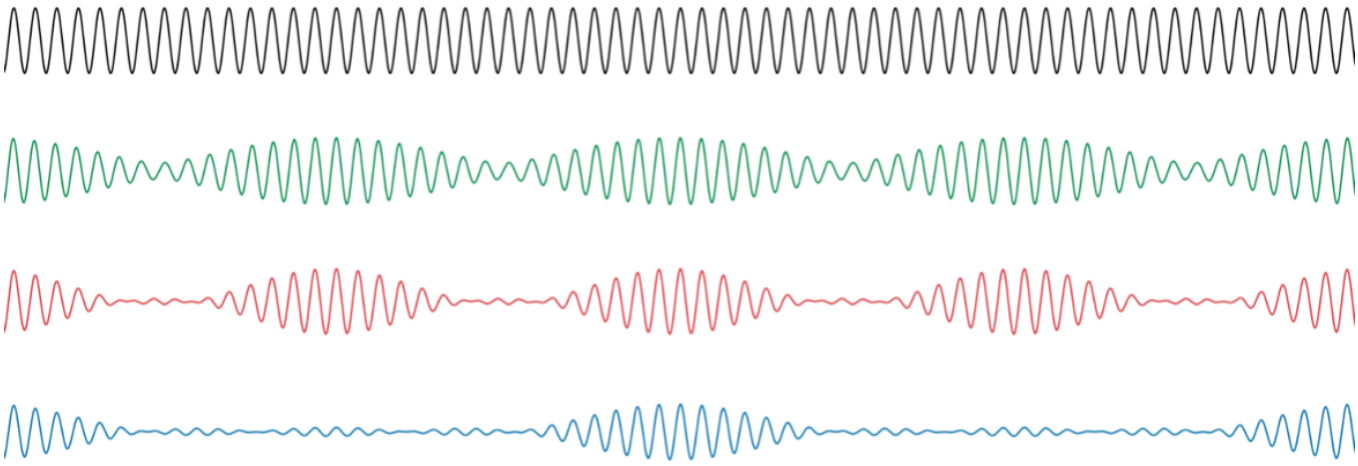


All we did here was add a third plane wave with a wave number *below* that of the original wave by the same amount as the second wave's wave number was larger:

$$\psi(x) = A\cos kx + 0.6A\cos(k+4\delta k)x + 0.6A\cos(k-4\delta k)x$$

Let's see what happens if we "fill in" a couple of the wave number gaps. That is, to get the above result, we added & subtracted $4\delta k$ to the wavenumber of the original wave for the added waves. Let's add & subtract $2\delta k$, and since it is closer to the original frequency, we'll weight its amplitude more as well ($0.8A$). The result is:

Figure 4.5.3 – Superposition of Five Plane Waves



The wave function for this fourth case is:

$$\psi(x) = A\cos kx + 0.8A\cos(k+2\delta k)x + 0.8A\cos(k-2\delta k)x + 0.6A\cos(k+4\delta k)x + 0.6A\cos(k-4\delta k)x$$

It should be clear what happens if we continue this program indefinitely – we are left with a single *wave packet*, as all the other bulges get pushed out to infinity. This localizes our free particle, giving it a finite probability of being found within a given region. This localization improved as we added the number of possible wavelengths (momenta) that could be measured. So we can improve our knowledge of the particle's position at the cost of knowing about its momentum, and vice-versa.

Spectral Content and Fourier (Again!)

Adding together lots of harmonic functions is exactly what we did when we did Fourier decomposition of periodic functions, but this is slightly different. Here we are creating a *non-periodic* function by adding together harmonic functions. The wave packet (in the limit of adding every wave number) is completely isolated – its bulge brethren are long gone – so it doesn't repeat and is therefore not periodic.

Notice that the bulges separated when we inserted plane waves with wave numbers *between* the ones we already had in place. In order to get the bulges separated to infinity we need to fill in *all* of those in-between wave numbers. Of course, there is a whole continuum of these available, so rather than use a sum of harmonics as we did with Fourier series for periodic waves, we need to use an integral to capture all of the harmonics. So to replace the Fourier series for periodic functions, we now have what is called the *Fourier transform* for non-periodic functions. We can see how to make the extension from the Fourier series to the Fourier transform by looking again at the Fourier series (Equation 1.7.5). The series is over the integer n, which ranges from $-\infty$ to $+\infty$, and this integer increments the wave number $k_n = \frac{2\pi n}{\lambda}$. If we now add over the continuum of wave numbers, this sum becomes an integral. The coefficients that are the amplitudes of the harmonic waves (the "recipe" of the wave being decomposed) depend upon the value of n in each case (or equivalently, the wave number), so the same is true in the continuous case. Even the sine and cosine are represented in the transform, though in a way that is somewhat different than in the series. The end result is:

$$\psi(x) = \int_{-\infty}^{+\infty} \left[A(k)\cos kx + iA(k)\sin kx\right] dk \tag{4.5.5}$$

The function $A(k)$ is the "recipe" for the transform, and yes it is the same function multiplying both the cosine and sine functions. Also, yes, the imaginary 'i' makes an appearance here. This gives us a more compact way to express the transform, using the Euler identity:

$$\psi(x) = \int_{-\infty}^{+\infty} A(k)\, e^{ikx}\, dk \tag{4.5.6}$$

In the case of the Fourier series, we had a means for computing the a_n and b_n coefficients, and the same is true for $A(k)$:

$$A(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \psi(x)\, e^{-ikx}\, dx \tag{4.5.7}$$

This is usually referred to as the *inverse Fourier transform*.

In Equation 4.4.10 we introduced the probability amplitude for measuring a particle's momentum (sometimes referred to as the wave function in "momentum space"). Given that the wave number is proportional to the momentum ($p = \frac{h}{\lambda} = \frac{h}{2\pi}\frac{2\pi}{\lambda} = \hbar k$), the "recipe" that gives the amount of each plane wave of a given wave number should be very closely related to the probability amplitude for momentum. Well, it turns out that these are simply proportional:

$$\phi(k) = \sqrt{2\pi}\, A(k) \tag{4.5.8}$$

This results in a nicely-symmetric relationship between the position and momentum probability densities:

$$\phi(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \psi(x)\, e^{-ikx}\, dx \tag{4.5.9}$$

$$\psi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \phi(k)\, e^{+ikx}\, dk \tag{4.5.10}$$

We will not show it here, but it is not difficult to prove that with this relationship, if either $\psi(x)$ or $\phi(k)$ is normalized, then so is its counterpart.

Digression: Dirac Brackets Again

There is a nice way to express all of this in terms of bras and kets. Thinking of them once again as vectors, we can put a general vector into any "basis" (set of unit vectors), by dotting the vector with those unit vectors to get its components, then multiplying by the unit vectors. For example:

$$\vec{v} = \hat{i}v_x + \hat{j}v_y + \hat{k}v_z = \hat{i}\left(\vec{v}\cdot\hat{i}\right) + \hat{j}\left(\vec{v}\cdot\hat{j}\right) + \hat{k}\left(\vec{v}\cdot\hat{k}\right)$$

In the bracket notation, there are an infinite number of these unit vectors, so we have to integrate to add them all up. Using the position "unit vectors" $|x\rangle$ and a state vector $|\psi\rangle$, we express the same vector decomposition as above this way:

$$|\psi\rangle = \int_{\text{all x}} |x\rangle \langle x|\psi\rangle \, dx$$

Now we can get the momentum wave function, which are components of the state vector using the momentum "unit vectors":

$$\phi(p) = \langle p|\psi\rangle = \int_{\text{all x}} \langle p|x\rangle \langle x|\psi\rangle \, dx$$

And now we get back the same relation as above if we have the dot product of the momentum and position unit vectors equal to:

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi}} e^{ikx}$$

The Heisenberg Uncertainty Principle

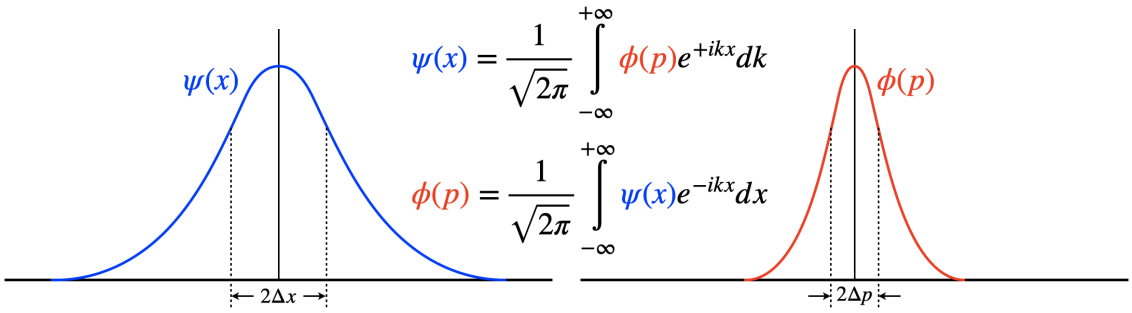
Now that it's quite clear that there is an inverse relationship between the uncertainty of measurements of position and momentum, we can state it formally as was first done by Werner Heisenberg. The specific predicament of a particular particle will define what the uncertainty will be in measuring either the position or the momentum, and we can change the conditions of our experiment to improve the certainty of our measurement of either of these quantities, but as we improve one, we necessarily worsen the other. According to Heisenberg's math, we get the following inequality expressing the principle that bears his name:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \tag{4.5.11}$$

No matter how we devise our experiment to measure x and p, when we compute their uncertainties statistically, we will find that the product of these uncertainties will always come out to a number no less than $\frac{\hbar}{2}$.

This principle is really well demonstrated through the Fourier transform. If we consider a localized (in position) wave packet (by "localized," we mean that the probability amplitude for measuring various positions drops off very rapidly far away from the center), and then Fourier-transform this function, we get the wave function of the same particle expressed in terms of its spectral content (probability amplitude for measuring various momenta).

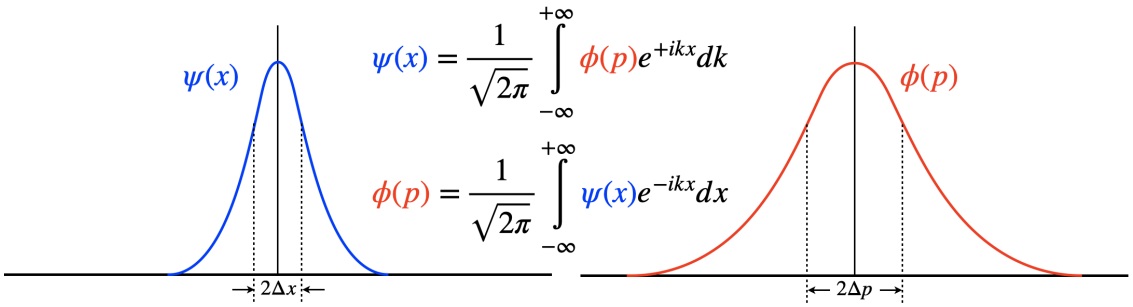
Figure 4.5.4 – Particle Wave Function Expressed in Terms of Position and Momentum



The uncertainties in position and momentum can be calculated in the usual way from the probability densities that come from these wave functions, and these are expressed in the diagram above.

Now suppose we change the physical conditions that brought about this quantum state. For example, suppose we change the way we take measurements so that we better-confine our knowledge of the position of the particle. This will serve to "tighten" its wave packet. When we take the Fourier transform of this new position wave packet, the momentum wave packet broadens.

Figure 4.5.5 – Same Particle with Position Measured More Precisely



Heisenberg's principle states that even if one provides the ideal conditions for the particle, this inverse relationship between the position and momentum uncertainties results in a limit to the minimum value that the product of these uncertainties can attain.

One last comment here: Notice that above we used the phrase "change the physical conditions" and "measure differently" interchangeably. This is a very important aspect of quantum theory. The probabilities we measure are dependent upon physical conditions, and the process of making measurements *necessarily affects these conditions*. We have encountered this once before, when we discussed watching electrons pass through a double slit, back in Section 3.5.

CHAPTER OVERVIEW

5: Matter Waves

5.1: The Schrödinger Wave Equation

Comparing Matter Waves and Light Waves

The fact that we get the same results for the double slit for both matter and photons tells us that matter wave functions must look the same (have the same mathematical form), and behave the same (satisfy superposition). Nevertheless, there must be *something* different about these two cases, as they are very different quanta, physically. We will see that while these quanta have wave functions of the same form, their physical differences lead to different wave equations! Let's begin our exploration with a look at a plane wave with a wavelength λ and frequency f, and we will represent it with a cosine function.

$$f(x,t) = A\cos\left(\frac{2\pi}{\lambda}x \pm 2\pi f t\right) \tag{5.1.1}$$

We can make the link between the physical properties of photons and the wave function explicit by using Planck's and de Broglie's relations:

$$E = hf \text{ and } p = \frac{h}{\lambda} \quad \Rightarrow \quad A\cos\left(\frac{2\pi}{\lambda}x \pm 2\pi f t\right) = A\cos\left(\frac{2\pi p}{h}x \pm \frac{2\pi E}{h}t\right) = A\cos\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) \tag{5.1.2}$$

If we plug this wave function into the wave equation that we know so well, we get confirmation that it works for light:

$$\frac{\partial^2}{\partial x^2}A\cos\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}A\cos\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) \quad \Rightarrow \quad -\frac{p^2}{\hbar^2}A\cos\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) = -\frac{1}{c^2}\frac{E^2}{\hbar^2}A\cos \tag{5.1.3}$$

$$\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) \quad \Rightarrow \quad E = pc$$

We want to use the same wave function for matter, but this wave *equation* won't work, because it does not yield the proper relationship between energy and momentum for particles with mass.

Schrödinger's Equation for Free Particles

We technically should find a wave equation for matter that satisfies the energy/momentum relation for relativity, but this turns out to be tougher to do mathematically, and historically this was not done, either. Instead, we'll assume that the particle is moving at a speed that is not relativistic, and we'll use the Physics 9HA-level relationship between kinetic energy and momentum:

$$KE = \frac{p^2}{2m} \tag{5.1.4}$$

For a freely-moving electron (we'll deal with electrons under the influence of forces later), we need a wave equation that gives us the correct relation between energy and momentum, but still gives us a harmonic wave solution, from which we can build more general waves, and that interferes in the same way that a light wave does. Notice that if our wave function has the same coefficients for x and t as for light, then we need two derivatives with respect to x (to give us the p^2), but only one with respect to t (so that we get only one factor of E). Also, each derivative brings out a factor of $\hbar^{-1}$, so we need to multiply by a factor of $\hbar$ for each derivative. We no longer require the $\frac{1}{c^2}$ factor on the right side of the equation, but we do need a factor of $\frac{1}{2m}$ on the right hand side, to construct the kinetic energy/momentum relation. So let's try this:

$$\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) = \hbar\frac{\partial}{\partial t}\psi(x,t) \tag{5.1.5}$$

Does this produce a harmonic wave function like the one in Equation 5.1.2? The constants that come from the chain rule all work out nicely, but this wave equation falls short in the derivative itself – the single derivative of cosine on the right gives a (negative) sine function, which doesn't match the cosine that comes from two derivatives on the left side.

A guy named Erwin Schrödinger didn't give up when he got this close. He realized that just as light waves have two parts (electric and magnetic), so too should matter waves. Here's how he incorporated two parts to the wave function: He allowed it to be a *complex number*. The real part of the wave function would be one part of the matter wave, and the imaginary part another. And just like for EM waves where changing electric fields give rise to magnetic fields and vice-versa, the real and imaginary parts of this wave function also mix. His solution is now known as *Schrödinger's equation* (for a free particle):

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) = i\hbar\frac{\partial}{\partial t}\psi(x,t) \tag{5.1.6}$$

Harmonic (plane wave) solutions to this differential equation look like:

$$\psi(x,t)=\psi_o\left[\cos\left(\frac{p}{\hbar}x\pm\frac{E}{\hbar}t\right)+i\sin\left(\frac{p}{\hbar}x\pm\frac{E}{\hbar}t\right)\right] \tag{5.1.7}$$

We can shorten the formula for the wave function using the Euler identity:

$$e^{i\theta}=\cos\theta\pm i\sin\theta \quad\Rightarrow\quad \psi(x,t)=\psi_o e^{i\left(\frac{p}{\hbar}x\pm\frac{E}{\hbar}t\right)} \tag{5.1.8}$$

As we saw in Section 4.5, a free particle does not automatically have a well-defined momentum. That is, while we have a plane wave solution to the free particle Schrödinger equation, that doesn't mean it is the only solution. Linear combinations of these plane waves can produce localized particles, the momentum (or kinetic energy) of which can be measured to be a wide range of values.

Wave Functions are Complex-Valued

Thanks to the work of Schrödinger, we now know that whatever it is that is "waving" in matter waves (we still don't have a good answer to that!), it is complex-valued. This leads to some complications that we have alluded-to before. first and foremost, we need the probability density to be positive-definite, and until now we have thought of the probability density as being the "square" of the probability amplitude, just like any wave's intensity is proportional to the square of its amplitude. But the square of a complex number is not positive in general, nor, for that matter, is it even a real number!

The solution to this is to insist that the probability density is the *magnitude-squared* of the probability amplitude (the wave function). The "magnitude-squared" operation is achieved by multiplying a complex number by its complex conjugate (the same number with the signs of all the imaginary i's switched):

$$\mathcal{P}(x)=|\psi(x)|^2=\psi^*(x)\,\psi(x) \tag{5.1.9}$$

An interesting consequence of this is that it leaves the wave function ambiguous. No matter how many boundary conditions we account for, we can never obtain an exact function for $\psi(x)$. In particular, we can always multiply it by a constant complex number with a magnitude of 1. This is because we can't *observe* the wave function, we can only see its probabilistic consequences, and these are unchanged by this *arbitrary quantum phase*:

$$\psi_2(x)=e^{i\delta}\psi_1(x) \quad\Rightarrow\quad |\psi_2(x)|^2=\left(e^{i\delta}\psi_1(x)\right)^*\left(e^{i\delta}\psi_1(x)\right)=\left(e^{-i\delta}\psi_1^*(x)\right)\left(e^{i\delta}\psi_1(x)\right)=e^0\psi_1^*(x)\,\psi_1(x) \tag{5.1.10}$$
$$=|\psi_1(x)|^2$$

Given that both of these wave functions $\psi_1(x)$ and $\psi_2(x)$ produce the same probability density, both will predict exactly the same probabilities of experimental results, and are therefore equally valid.

At this point it should also be noted how the complex-valued wave function is related to the Dirac bracket notation. When the bracket is closed, we have a complex number. If we reverse the order of the bracket, the value is changed to the complex conjugate:

$$\langle x|\psi\rangle=\psi(x) \quad \langle\psi|x\rangle=\psi^*(x) \tag{5.1.11}$$

If we wish to take the dot product of two state vectors $|\psi_1\rangle$ and $|\psi_2\rangle$ (this tells us how much they "overlap"), which we write most simply as $\langle\psi_1|\psi_2\rangle$, we can express this in terms of their wave functions. As with any dot product, it equals the sum (integral) of the products of their components (wave functions):

$$\langle\psi_1|\psi_2\rangle=\int_{-\infty}^{+\infty}\langle\psi_1|x\rangle\langle x|\psi_2\rangle\,dx=\int_{-\infty}^{+\infty}\psi_1^*(x)\,\psi_2(x)\,dx \tag{5.1.12}$$

What if the Particle is not "Free"?

It would be awfully boring if all we studied was particles that didn't interact with anything. So what does Schrödinger have to say about particles that experience forces? It turns out that the extension to such particles is a simple one. We have been assuming that the energy that appears in the argument of the harmonic wave function (Equation 5.1.2) is the kinetic energy, but what if it is just the *total* energy? Naturally these two values are one and the same in the case of a free particle, but a particle under the influence of a force also has some potential energy. So with this understanding that it is the total the energy of a particle that is cataloged in the wave function, we simply add a term to the left side of the Schrödinger equation, to go along with the kinetic energy term that is already there. Calling the potential energy (which depends only upon the position of the particle) $V(x)$, we have as the full Schrödinger equation:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t)+V(x)\psi(x,t)=i\hbar\frac{\partial}{\partial t}\psi(x,t) \tag{5.1.13}$$

The free particle case is obviously recovered when the potential energy is a constant zero value. We will spend the bulk of our remaining time deriving consequences for some especially-instructive functions for $V(x)$.

5.2: States of Definite Energy

Separation of Variables

Now that we have the Schrödinger equation, we can link the physical circumstances the particle finds itself in to its wave function, and from that we can make probabilistic predictions of its behavior. So now our attention turns to actually *solving* this differential equation. Like the classical wave equation before it, the Schrödinger equation in one dimension (which we will focus on for quite awhile) is a partial differential equation with two variables. As we did with string waves all the way back in Section 1.2, we will start by looking only at a certain family of solutions to this differential equations – those that can be separated into a product of two functions of single variables. While to this point we have used the symbol ψ for both the time-dependent wave function $\psi(x,t)$ and the wave form $\psi(x)$, from this point on to avoid confusion, we will use the capitalized version Ψ for the time-dependent wave function $\Psi(x,t)$ and retain the lower-case ψ for the time-independent version.

Following the separation method we used before, we write the full wave function as the product of two others:

$$\Psi(x,t)=\psi(x)\cdot\tau(t) \tag{5.2.1}$$

This time we plug it into the Schrödinger equation:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t)+V(x)\Psi(x,t)=i\hbar\frac{\partial}{\partial t}\Psi(x,t) \quad\Rightarrow\quad -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}[\psi(x)\cdot\tau(t)]+V(x)[\psi(x)\cdot\tau(t)] \tag{5.2.2}$$
$$=i\hbar\frac{\partial}{\partial t}[\psi(x)\cdot\tau(t)]$$

The partial derivatives only act on the function of the same variable, so representing derivatives of the single variable in each case with primes, we have:

$$-\frac{\hbar^2}{2m}\psi''(x)\tau(t)+V(x)\psi(x)\tau(t)=i\hbar\psi(x)\tau'(t) \tag{5.2.3}$$

Dividing both sides of the equation by $\psi(x)\tau(t)$ gives:

$$\frac{-\frac{\hbar^2}{2m}\psi''(x)+V(x)\psi(x)}{\psi(x)}=i\hbar\frac{\tau'(t)}{\tau(t)} \tag{5.2.4}$$

The left side of this equation is exclusively a function of x, while the right side is exclusively a function of t. For these to be equal for all values of x and t, they must equal a common constant., and that gives us two ordinary (single variable) differential equations that share a common constant. Before we continue with the math, we should think about what physical quantity this constant might be. When we derived the Schrödinger equation, we noted that the first term links the wave function to the particle's kinetic energy, and of course the second term links it to the potential energy. Given that this sum will equal a constant multiplied by the wave function, it seems reasonable to conclude that the constant is the total energy, E. This also works with the right side of the equation, given the relationship we have seen for the time portion of the plane wave:

$$\frac{-\frac{\hbar^2}{2m}\psi''(x)+V(x)\psi(x)}{\psi(x)}=E=i\hbar\frac{\tau'(t)}{\tau(t)} \tag{5.2.5}$$

$$\Rightarrow\quad -\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x)+V(x)\psi(x)=E\psi(x)\ ,\quad i\hbar\frac{d}{dt}\tau(t)=E\tau(t) \tag{5.2.6}$$

The time-independent wave function that satisfies the first differential equation depends upon the specifics of $V(x)$ (and of course boundary conditions), but the form of the solution to the second differential equation (absent of an initial condition) looks like:

$$\tau(t)=\tau_o e^{-i\omega t}\ ,\quad \text{where: } \omega\equiv\frac{E}{\hbar} \tag{5.2.7}$$

Ultimately we will use the functions $\psi(x)$ and $\tau(t)$ to rebuild $\Psi(x,t)$, and of course $\Psi(x,t)$ will need to be normalized in order for the probability density derived from it to give proper results. The probability density will then be:

$$\mathcal{P}(x,t)=|\Psi(x,t)|^2=\psi^*(x)\tau^*(t)\psi(x)\tau(t)=|\psi(x)|^2|\tau_o|^2e^{+i\omega t}e^{-i\omega t}=|\tau_o|^2|\psi(x)|^2 \tag{5.2.8}$$

It will be easier for us later if we insist that the time-independent wave function $\psi(x)$ itself be normalized, and with the total probability density also normalized, we choose the simplification that $\tau_o=1$, and the time portion of our separated wave function is simply equal to $e^{-i\omega t}$.

Special Physical Conditions

There are two special physical (measurable) conditions that accompany these separable solutions to the Schrödinger equation. The first we can see from the previous equation, which (with our choice of τ_o) reduces to simply:

$$\mathcal{P}(x,t) = |\psi(x)|^2 \tag{5.2.9}$$

This shows that in fact the probability density does not change with time. Physically this means that it when we compute the probability of finding the particle at a specific location, that probability doesn't change with time. We have seen this many times already – for example, the probability of a particle landing at a specific position on a screen after passing through a double-slit does not fluctuate with time. For this reason, we call such solutions the *stationary-state* solutions to Schrödinger's equation – the quantum state remains unchanging ("stationary") in time. It should be emphasized that the wave function itself *is* changing with the passage of time – its quantum phase is oscillating with frequency ω. But as we have discussed previously, the wave function does not determine probabilities directly – only its *magnitude-squared* has any significance (it is the probability density), and for these separated wave functions the time part of the wave function – while it exists – does not contribute.

The second physical condition that accompanies this type of solution to the Schrödinger equation is the energy. As we know, whenever we measure a quantity for a quantum state, in general that quantity can take on several values. A wave packet has many measurable positions and momenta, for example. But with a single energy emerging from the separation of variables, we see that measurement of the energy of a particle described by such wave function can only yield a single value.

So stationary quantum states are states of definite energy. It is useful to introduce language that we will use frequently from now on. We say that the energy associated with one of these wave functions is the *eigenvalue* of energy ("*eigen*" is German for "own" - this is the value of energy "owned" by this state). Such a state is also referred to as an *eigenstate* of energy. Plane wave states (which have single values of momentum) are eigenstates of momentum and, as it happens, of energy as well, since they have a fixed kinetic energy and no potential energy. It is therefore not surprising that plane wave functions are separable.

Building General Wave Functions

When we first used separation of variables, we found that the separated single-variable functions were harmonic, and that these solutions could be combined together in a Fourier series to construct any periodic function. If the potential energy of the particle is zero (i.e. it is free), then the separated functions are also harmonic (plane waves), and those too can be combined to make more general free particle states (wave packets). Well it turns out that even more generally, separated solutions to the Schrödinger equation with a potential energy can also be put together to construct any general state subject to that potential *even though the separated solutions are not strictly harmonic* when the potential is not zero (or a constant). This fact is the statement of what is called the *spectral theorem*, but it is well beyond the scope of this course to go any further than to state this as fact. The technical language you will hear in this regard is, "the collection of the stationary state solutions for the Schrödinger equation with a given potential energy forms a complete set of states". *Completeness* refers to the fact that all of the possible solutions can be constructed in this way.

Another way to think of this is once again in terms of vectors. The three unit vectors $\hat{i}$, $\hat{j}$, and $\hat{k}$ are "complete" in that any 3-dimensional vector can be formed from a linear combination of these. So the various separable solutions form a (typically infinite) set of orthogonal functions, linear combinations of which can be constructed to build any general solution.

It's important to understand that when using multiple stationary state solutions of the Schrödinger equation to build a more general solution, the result is *not* a stationary state solution. When we used two plane waves to construct another solution to the free particle Schrödinger equation, we did not get back another plane wave solution (we can't put together two harmonic waves with different wavelengths to get a new harmonic wave with a new wavelength!). To see this explicitly, let's consider two stationary-state solutions $\Psi_1(x,t)$ and $\Psi_2(x,t)$, associated with energies E_1 and E_2. If we create a new solution (which in this case is not normalized, but we'll come back to that issue later) by mixing these in equal amounts, we get:

$$\Psi_{tot}(x,t) = \Psi_1(x,t) + \Psi_2(x,t) = \psi_1(x)\,e^{-i\omega_1 t} + \psi_2(x)\,e^{-i\omega_2 t} \tag{5.2.10}$$

$$\Rightarrow \quad \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right]\Psi_{tot}(x,t) = \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right]\psi_1(x)\,e^{-i\omega_1 t} + \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right]\psi_2(x)\,e^{-i\omega_2 t}$$

$$= E_1\psi_1(x)\,e^{-i\omega_1 t} + E_2\psi_2(x)\,e^{-i\omega_2 t}$$

Clearly this does not have the form of a stationary state solution, as the functions of time cannot be made to go away.

5.3: Operators and Observables

Quantum State Information

Something we discussed only obliquely in an earlier section is the idea of a quantum state and the information contained within it. There are some very strange features associated with the concept of a quantum state. High among those it the fact that it is *non-local*. We are used to the classical notion that the mass, charge, and other features of a particle are located *at the particle* – we can literally point to the point to the position in space where these quantities can be found. But now we have to accept the fact that even the location of the particle itself is not something that is well-defined. The wave function of a particle exists everywhere in space at the same time, and it isn't until it interacts with a measuring device that its location is defined. To emphasize this point: It isn't that the particle is somewhere but we just don't know where (like the result of a coin flip still concealed by someone's hand), it actually not located anywhere until it is observed.

All of this mysterious mumbo-jumbo might tempt us to throw up our hands in despair that we can't do any of the predictive science that we've become accustomed to in classical physics, but the quantum state of a particle *does* contain useful information about it. Indeed, the theory claims that the quantum state contains *all* of the accessible information about the particle. Much of it is probabilistic, but this is still useful. We have already discussed a bit about how to extract this information from the quantum state – we take averages using the probability density. We have slightly oversimplified this process, but we will correct that now.

Expectation Value Computation

The key to pulling information from the quantum state is calculating expectation values. Even when we want to compute uncertainties, to do this we need to be able to compute averages. So far, we have seen that the method for doing this is to multiply the quantities measured by their associated probability densities, and integrate over all the possible values. For example, if we wish to calculate the average position:

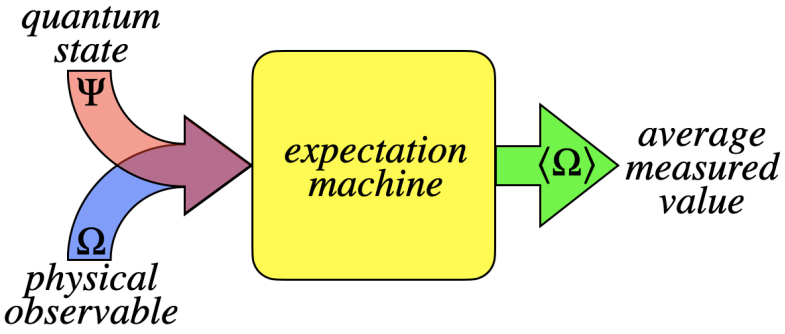
$$\langle x \rangle = \int_{-\infty}^{+\infty} \mathcal{P}(x)\; x\; dx = \int_{-\infty}^{+\infty} |\Psi|^2\; x\; dx = \int_{-\infty}^{+\infty} \Psi^* \Psi\; x\; dx \tag{5.3.1}$$

If we instead which to calculate the average momentum, we can Fourier-transform the position wave function to get the momentum version, and use it in the integral along with a k and dk in place of x and dx (this will give an average wave number, which can then be multiplied by $\hbar$ to get an average momentum). Notice that in the momentum case we can't use the usual $\mathcal{P}(x)$, because the momentum values are not a function of x. Lucky that we have the Fourier transform! But what if there are other observable quantities for which we wish to compute an average (energy, angular momentum, etc.)?

Quantum mechanics provides an alternate means that is totally equivalent to the one above for position and momentum, without the need for a Fourier transform, and which works for other quantities. What is more, this process for computing averages embodies the idea that measurements affect the very quantum state they seek to measure. We'll start with a basic description for how this works...

We begin with two things: The quantum state we are working with, and the observable whose the expectation value we wish to compute. We throw these both into a machine, which in turn spits out the expectation value:

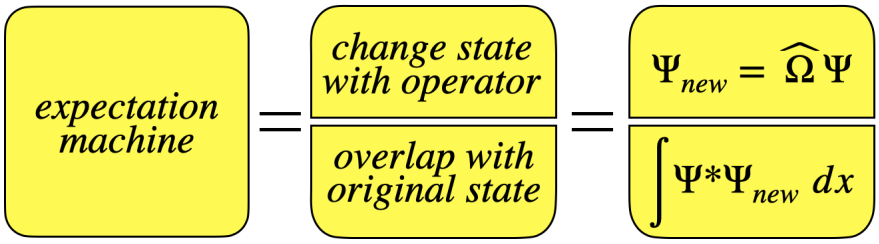
Figure 5.3.1 – Expectation Value Machine



Well, this is just fine, but of course we need to peek behind the curtain to see precisely how this expectation machine functions. It works in a few steps:

1. It invents an *operator* that belongs to the given observable. It is one of the postulates of quantum theory that every quantity that can be measured and is stored in a quantum state has an associated operator.
2. The operator "acts upon" the state, changing it into a new state. This is the part of the process where the observation of a physical property of a particle alters the state of the particle being observed.
3. The "overlap integral" of new state and the original state is computed. As we have said before, the integral of the product of two functions is like a dot product (e.g. odd functions are "orthogonal" to even functions, as their overlap integral is zero). So this overlap integral gives us a sense of how far the wave function has been altered from its original one.

Figure 5.3.2 – Machine Inner Workings



It's probably not immediately clear how this process gives us the average value we wish to compute, so let's look a bit closer.

The Position and Momentum Operators

Let's look first at the simple case of $\langle x \rangle$. In this case, the "new state" has a wave function for position that is simply the product of x and the previous wave function:

$$\Psi_{new} = x\,\Psi \quad \Rightarrow \quad \langle x \rangle = \int_{-\infty}^{+\infty} \Psi^* \Psi_{new} dx = \int_{-\infty}^{+\infty} \Psi^* \left(x\,\Psi\right) dx = \int_{-\infty}^{+\infty} x\,|\Psi|^2 dx = \int_{-\infty}^{+\infty} x\,\mathcal{P}(x)\,dx \tag{5.3.2}$$

If we now wish to do the same with momentum, it is not clear how the momentum operator creates a new quantum state from the old one, when we describe that quantum state in terms of position. We *do* know how it changes the quantum state when it is described in terms of momentum (or wave number) – it works the same way as x did:

$$\Phi_{new} = p\,\Phi \quad \Rightarrow \quad \langle p \rangle = \int_{-\infty}^{+\infty} \Phi^* \Phi_{new} dk = \int_{-\infty}^{+\infty} \Phi^* \left(p\,\Phi\right) dk = \int_{-\infty}^{+\infty} p\,|\Phi|^2 dk = \int_{-\infty}^{+\infty} p\,\mathcal{P}(k)\,dk \tag{5.3.3}$$

But now we are interested in how the momentum affects the quantum state *when the wave function is viewed in terms of position*. To do this, we turn to our "translation" device - the Fourier transform. Noting that $\Phi_{new} = p\Phi$, we can do an inverse Fourier transform to get both the original wave function Ψ and the newly-altered function Ψ_{new}:

$$\Psi = \int_{-\infty}^{+\infty} \Phi\, e^{ikx} dk\,, \quad \Psi_{new} = \int_{-\infty}^{+\infty} \Phi_{new}\, e^{ikx} dk = \int_{-\infty}^{+\infty} p\,\Phi\, e^{ikx} dk = \int_{-\infty}^{+\infty} \hbar k\,\Phi\, e^{ikx} dk \tag{5.3.4}$$

Now we seek some operation we can perform on Ψ that can give us Ψ_{new}. Without further ado, we declare that if we act on Ψ with the operation $-i\hbar\frac{d}{dx}$, that will do the trick. The wave function Φ is only a function of k (not x), so:

$$\Psi_{new} = -i\hbar\frac{d}{dx}\Psi = -i\hbar\frac{d}{dx}\int_{-\infty}^{+\infty} \Phi\, e^{ikx} dk = -i\hbar\int_{-\infty}^{+\infty} \Phi\frac{d}{dx} e^{ikx} dk = \int_{-\infty}^{+\infty} \hbar k\,\Phi\, e^{ikx} dk \tag{5.3.5}$$

To summarize, the x-direction momentum operator for use on wave functions expressed in terms of position is (when we eventually go beyond 1-dimension, this will become a partial derivative):

$$\hat{p}_x = -i\hbar\frac{d}{dx} \tag{5.3.6}$$

The little "hat" above the p is a reminder to us that we are talking about an operator that changes a quantum state, and not just the value of momentum. What this operator actually is depends upon the type of wave function it is acting on. That is:

$$\hat{p}_x \, \psi(x) = -i\hbar \frac{d}{dx} \psi(x) \; , \quad \hat{p}_x \, \phi(k) = \hbar k \, \phi(k) \tag{5.3.7}$$

Similarly, the operator $\widehat{x}$ is just the function $f(x) = x$ when acting on a wave function expressed in terms of position, and will involve a derivative when acting on a wave function expressed in terms of wave number (it is left as an exercise o the reader to determine the operator $\widehat{x}$ that acts on $\phi(k)$).

Going back to the original discussion of computing expectation values, we see that we have:

$$\langle p \rangle = \int_{-\infty}^{+\infty} \Psi^* \left(\hat{p} \; \Psi \right) dx = \int_{-\infty}^{+\infty} \Psi^* \left(-i\hbar \frac{d}{dx} \Psi \right) dx \tag{5.3.8}$$

Building More Operators

We can build new operators for other physical observables from $\widehat{x}$ and $\hat{p}$. Most notable among these is the kinetic energy operator:

$$\widehat{KE} = \frac{\hat{p}^2}{2m} = \frac{1}{2m} \left(-i\hbar \frac{d}{dx} \right) \left(-i\hbar \frac{d}{dx} \right) = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \tag{5.3.9}$$

This looks familiar! It is precisely what acts on the wave function in the Schrödinger equation, which we already said accounts for the particle's kinetic energy. Now we see the Schrödinger equation in a whole new light – as an equation that relates the effects of operators. The potential $V(x)$ is just a function of x, so it is an operator formed from $\widehat{x}$. Together, the operators $\widehat{KE}$ and $\widehat{V(x)}$ account for the total energy, and as a shorthand we sometimes use:

$$\widehat{H} \equiv \widehat{KE} + \widehat{V(x)} \tag{5.3.10}$$

This "total energy operator" is commonly referred to as the *Hamiltonian*. Note that Schrödinger's equation states that the Hamiltonian's actions in on the wave function expressed in terms of position are equivalent to another operator's actions. The other operator (sometimes called the "total energy operator") is what we see on the right hand side of the Schrödinger equation:

$$\widehat{E} \equiv i\hbar \frac{\partial}{\partial t} \tag{5.3.11}$$

Uncertainty Principle

We have already seen that measurements of position and momentum are "incompatible" in that the measurement of one affects the measurement of the other – the more we take care to precisely one of them, the less able we are to measure the other. This comes through very clearly with this idea that operators change quantum states into new states. We would expect that the alteration of the state by one of these two operators will have an effect on the measurement of the expectation for the other, and it does. Suppose, for example, that for whatever reason, we wish to know the expectation value of the product of the position and momentum, xp. We follow our "expectation machine" method, but since we now have two operators, we have to do them in sequence – first change the quantum state by one of the operators, and then by the other. If we use the momentum operator first, we get:

$$\Psi_{new} = \widehat{x} \; \left(\hat{p} \; \Psi \right) = x \; \left(-i\hbar \frac{d}{dx} \Psi \right) = -i\hbar x \; \frac{d\Psi}{dx} \tag{5.3.12}$$

But if we perform the operation in the other order, we get a different result:

$$\Psi_{new} = \hat{p} \; \left(\widehat{x} \; \Psi \right) = -i\hbar \frac{d}{dx} (x\Psi) = -i\hbar \left(\Psi + x \; \frac{d\Psi}{dx} \right) \tag{5.3.13}$$

This effect of two operators "tripping over each other" is directly related to an uncertainty principle between those two operators – changing the state by one of them affects the measurement of the other. When two operators do not acheive the same result when performed in either order, we say that that do not *commute* with each other. Note that any function of $\widehat{x}$ (like $\widehat{V}(x)$) will commute with any other function of $\widehat{x}$, and any function of $\hat{p}$ (like $\widehat{KE}$) will commute with any other function of $\hat{p}$. So measuring the momentum will not have an effect on measuring the kinetic energy.

5.4: Eigenstates and Eigenvalues

Observable Values

Suppose we start with a quantum state that provides for a broad spectrum of measurements of some quantity, such as energy. What happens to that quantum state after we observe it? Well, the cat is out of the bag, in that now we know precisely the state's energy – the state doesn't go back to a probabilistic one unless we prepare it that way. A classical analog to this is the roll of a die. We have a die in a cup, shake the cup, and turn it over onto a tabletop. We don't know the roll of the die until we "measure" it by lifting the cup, but once we do, the die remains in that state unless we shake it in the cup again.

One thing we can say about the state of the die, even before we measure it, is that *when* we measure it, we are guaranteed to get one of the six possible outcomes. We compute the average roll of the die to be 3.5, and even though we call this the expectation value, we certainly don't ever expect to see 3.5 dots on the die staring up at us when we lift the cup.

The same applies to physical observables. It may be that the possible measurements lie on a continuum, making every outcome possible, but it may also be that only certain outcomes are possible (in a double slit experiment, positions at the dark fringes are not among the possible outcomes, for example). Once the physical quantity is measured, then the quantum state changes from a probabilistic description to a specific one, and that can only be one of the ones that was "allowed" by the physical situation.

There is one other thing we should say about observable values in quantum theory before moving on. One thing we have had to accept in our mathematical treatment of quantum theory is the presence of complex numbers. These unavoidably come into play whenever the phase of a wave function is important (namely, when there is interference). But measurements of real, physical quantities can never result in a number with an imaginary part. We should never see that a particle's energy is something like "$(3.0+2.2i)\,eV$"! Furthermore, we should never see an expectation value with an imaginary portion either. After all, besides calculating expectation values, we can get them by making lots of observations of the same state and averaging the numbers. If none of the numbers being averaged can have an imaginary part, then neither can their average.

Special Quantum States

These quantum states that exist after we make a measurement of a physical observable have the property that future measurements of that observable give the same result every time. This means that the expectation value of that observable is exactly that value, and the uncertainty is zero. Let's see what this means mathematically. Let's define the wave function $\psi_1(x)$ to be the state that always produces the same observable value ω_1, and we'll call the operator for that observable Ω. Then, using our "expectation machine" from the previous section, we have:

$$\langle \omega \rangle = \omega_1 = \int_{-\infty}^{+\infty} \psi_1^*(x) \left[\widehat{\Omega}\,\psi_1(x)\right] dx \tag{5.4.1}$$

The question is, in what way does the operator $\widehat{\Omega}$ alter the wave function $\psi_1(x)$? We will state without yet proving that it changes it to a new wave function that differs from the original only in that the original is multiplied by a constant real number. Thinking of the quantum state as a vector, this means that the vector is rescaled, but not rotated. We can see that the constant real number is simply ω_1:

$$\int_{-\infty}^{+\infty} \psi_1^*(x) \left[\widehat{\Omega}\,\psi_1(x)\right] dx = \int_{-\infty}^{+\infty} \psi_1^*(x) \left[\omega_1\,\psi_1(x)\right] dx = \omega_1 \int_{-\infty}^{+\infty} \psi_1^*(x)\,\psi_1(x)\,dx = \omega_1 \tag{5.4.2}$$

The last equality comes about because the wave function is normalized.

In Section 5.2 we introduced the label of an eigenstate and eigenvalues in the context of states of definite energy and those energy values. We see here that this notion generalizes to any observable. It is common practice to distinguish *eigenfunctions* (the wave functions associated with eigenstates) from more general wave functions with a label that indicates which eigenvalue it is linked to, so in general we would write:

$$\widehat{\Omega}\,\psi_i(x) = \omega_i\,\psi_i(x) \tag{5.4.3}$$

Eigenstates are "Complete"

Given that there is an eigenstate associated every possible value of an observable, then it should come as no surprise that quantum states that are not eigenstates (i.e. they produce many possible outcomes with different probabilities) can be written as a linear combination of eigenstates. We have already seen this idea of "completeness" in the context of building more general states from energy eigenstates, which were found using separation of variables, but now we can state that no matter what *basis* we use, these eigenstates behave like "unit vectors" that allow us to build any quantum state vector.

If the possible observable values are *quantized* (i.e. only come in discrete units), then a general wave function is constructed from a linear combination of the eigenfunctions:

$$\psi(x) = C_1\,\psi_1(x) + C_2\,\psi_2(x) + \cdots = \sum_{\text{all } i} C_i \psi_i(x) \tag{5.4.4}$$

If, on the other hand, the possible observable values lie on a continuum, then the linear combination requires an integral. We have actually seen this already! We know that a plane wave solution to the free particle Schrödinger equation (e^{ikx}) has a definite momentum ($p = \hbar k$). Each eigenstate of momentum has its own value of k, and these lie on a continuum for the free particle. So a general wave function is a linear combination (integral) summed over all of these eigenfunctions, with the coefficients of each eigenfunction expressed as "$A(k)$", giving us Equation 4.5.6.

Eigenstates are "Orthogonal"

Our description of eigenstates as the "unit vectors" of quantum states does not end with being able to construct general vectors. They also satisfy an orthogonality condition, like the one we discussed in Section 1.6:

$$\int_{-\infty}^{+\infty} \psi_i^*(x)\,\psi_j(x)\,dx = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{5.4.5}$$

The value of 1 comes about when $i = j$ because the wave function is normalized.

Using this fact, we can show that the coefficients in the linear combination are in fact probability amplitudes associated with measuring each of the respective eigenvalues when an observation is made on a general quantum state:

$$1 = \int_{-\infty}^{+\infty} \psi^*(x)\,\psi(x)\,dx = \int_{-\infty}^{+\infty} \left[C_1\,\psi_1(x) + C_2\,\psi_2(x) + \ldots\right]^* \left[C_1\,\psi_1(x) + C_2\,\psi_2(x) + \ldots\right] dx \tag{5.4.6}$$

The integrals of the cross-terms all vanish thanks to the orthogonality condition, leaving:

$$1 = C_1^* C_1 + C_2^* C_2 + \cdots = |C_1|^2 + |C_2|^2 + \ldots \tag{5.4.7}$$

The quantity $|C_i|^2$ is the probability of a measurement of an observable from a general state resulting in the eigenvalue of the i^{th} state.

Now we can also show why our "expectation machine" works:

$$\begin{aligned} \langle\omega\rangle &= \int_{-\infty}^{+\infty} \psi^*(x)\left[\widehat{\Omega}\psi(x)\right] dx \\ &= \int_{-\infty}^{+\infty} \left[C_1^*\,\psi_1^*(x) + C_2^*\,\psi_2^*(x) + \ldots\right]\left[C_1\,\widehat{\Omega}\psi_1(x) + C_2\,\widehat{\Omega}\psi_2(x) + \ldots\right] dx \\ &= \int_{-\infty}^{+\infty} \left[C_1^*\,\psi_1^*(x) + C_2^*\,\psi_2^*(x) + \ldots\right]\left[C_1\,\omega_1\,\psi_1(x) + C_2\,\omega_2\,\psi_2(x) + \ldots\right] dx \\ &= |C_1|^2\omega_1 + |C_2|^2\omega_2 + \ldots \\ &= P_1\omega_1 + P_2\omega_2 + \ldots \end{aligned} \tag{5.4.8}$$

The "altered" state $\Psi_{new} = \widehat{\Omega}\Psi$ used un the expectation machine is just what comes from weighting every eigenstate in the original state's "recipe" by the amount of its associated eigenvalue.

Simple Examples

In the case of a free particle plane wave moving in the $+x$-direction, the full wave function is:

$$\Psi(x,t) = Ae^{i(kx-\omega t)} \tag{5.4.9}$$

We fully expect this to be an eigenstate of momentum, kinetic energy, and total energy, and it is:

$$\begin{aligned}
\hat{p}\Psi(x,t) &= -i\hbar\tfrac{\partial}{\partial x}Ae^{i(kx-\omega t)} = \hbar k\Psi(x,t) \\
\widehat{KE}\Psi(x,t) &= -\tfrac{\hbar^2}{2m}\tfrac{\partial^2}{\partial x^2}Ae^{i(kx-\omega t)} = \tfrac{\hbar^2 k^2}{2m}\Psi(x,t) \\
\widehat{E}\Psi(x,t) &= i\hbar\tfrac{\partial}{\partial t}Ae^{i(kx-\omega t)} = \hbar\omega\Psi(x,t)
\end{aligned} \tag{5.4.10}$$

Simultaneous Eigenstates

In the case of a free particle plane wave, we see that it is an eigenstate of many observables at the same time. We already know that an eigenstate of momentum cannot simultaneously be an eigenstate of position due to the uncertainty principle, so being an eigenstate of two different observables at the same time is certainly not guaranteed. The deciding factor is something we mentioned at the end of Section 5.3. If the operators associated with two observables commute with each other – if the altered quantum state that results the consecutive actions of the operators is the same regardless of the order in which they are applied – then these two observables can share eigenstates.

CHAPTER OVERVIEW

6: One-Dimensional Models

6.1: Particle-in-a-Box, Part 1

Bound States

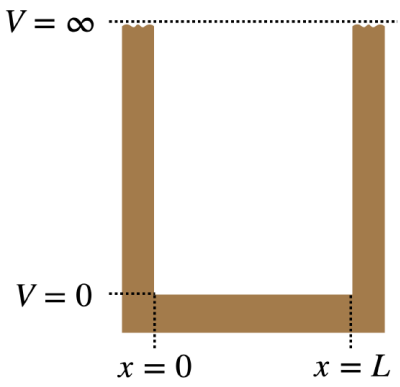
We have discussed at length the case of a free particle, and how we can construct general solutions from plane wave solutions to the Schrödinger equation, but now it's time to have a look at cases where particles are bound to some region by a force. We are staying in one dimension, so this force will need to act in both directions, always acting to keep the particle from straying too far from a central point. [In Physics 9HA, we called such a force a "restoring force."] It is unclear how to use the concept of a force when discussing the effect it has on entities that behave like waves, but since the Schrödinger equation accounts for a potential energy, we can certainly use that. As always, we wish to start as simply as possible, and build our way up to the more complicated cases. As we do this "build-up", we will try to sort out what features of bound states appear to be universal, and what features are special to the model we are examining.

The Infinite Square Well Potential

The simplest conceivable potential well allows us to keep most of the features of the free particle, but simply confines it between two impenetrable potential "walls." We will place these walls at $x = 0$ and $x = L$, and make them such that it is impossible for a particle of any finite energy to escape. The full mathematical description is:

$$V(x) = \begin{cases} 0 & 0 \le x \le L \\ \infty & x < 0,\ x > L \end{cases} \tag{6.1.1}$$

Figure 6.1.1 – Infinite Square Well



When we put this into the Schrödinger equation, we find that the wave function splits up easily into two parts: The part that is inside the well (where $V = 0$) which is simply the free particle equation (where the free particle can be traveling in either direction), and the part that is outside the well, which can only satisfy the Schrödinger equation if $\Psi(x,t)$ is identically zero. These two conditions sound very familiar – a wave that can be constructed from harmonic functions (like the free particle plane waves) and has endpoints that must remain fixed at zero – the wave function created by this potential should be similar to a standing wave on a string!

We should also say a word about the classical analog of this potential. Clearly the vertical potential wall corresponds to providing an infinite force, since $F = -\frac{dV}{dx}$. This is exactly what we would assume classically for a rigid ball colliding elastically with a rigid wall – the ball's momentum reverses direction instantly (and keeps the same magnitude), and since this requires a finite net impulse over an infinitesimally-short time period, the force must be infinite. We will come back to classical analogs like this occasionally throughout our study of bound states, to see how the quantum versions differ, and particularly to see how they converge at macroscopic scales.

Stationary-State Solutions

We now follow our prescribed program for finding wave functions from Schrödinger's equation, beginning with the separated stationary-state solutions. We are seeking the wave functions that satisfy:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + V(x) = E\psi(x) \tag{6.1.2}$$

Once we have the possible values of E (called the *energy spectrum*), we can use them to compute the oscillation frequencies $\omega = \frac{E}{\hbar}$, and then construct any general wave function solution for this potential by making linear combinations of the $\psi(x)$'s and their corresponding $e^{-i\omega t}$'s. So for the stationary-state wave functions, we essentially have a differential equation for each region (inside and outside the well):

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + \left\{ \begin{array}{ll} 0 & \text{(inside well)} \\ \infty & \text{(outside well)} \end{array} \right\} \psi(x) = E\psi(x) \tag{6.1.3}$$

Let's do the easy part first – outside the well. In this case, we see that an infinite number multiplies the probability amplitude $\psi(x)$ on the left side of the equation and a finite number multiplies it on the right. The only place where the second derivative cannot have any infinite effect on this equation over the entire outside region, so the only way this can be solved is for $\psi(x)$ to be identically zero outside the well. This also makes sense from a probability perspective – we would expect to *never* see the particle in the region outside the well, so we would expect this probability to be zero, which means we expect its probability amplitude to also be zero.

The solution for inside the well is not much tougher than outside, as it is the same differential equation that we had for the free particle. Stationary-state solutions consist of plane waves, which can be traveling in either direction. As the Schrödinger equation only takes into account energy, it doesn't select one direction over another, and the general stationary-state solution is a linear combination of both:

$$\psi(x) = Ae^{+ikx} + Be^{-ikx} \tag{6.1.4}$$

Now we have to apply the boundary conditions. The wave function *must* be continuous everywhere, most notably at the walls $x = 0$ and $x = L$. Since $\psi(x)$ vanishes just on the other side of the walls, we have that $\psi(0) = \psi(L) = 0$. Plugging this in gives:

$$\left. \begin{array}{l} \psi(0) = 0 = A + B \\ \psi(L) = 0 = Ae^{+ikL} + Be^{-ikxL} \end{array} \right\} \quad \Rightarrow \quad e^{+2ikL} = 1 \quad \Rightarrow \quad k_n = \frac{n\pi}{L}, \;\; n = 0,\, 1,\, 2, \ldots \tag{6.1.5}$$

We have subscripted the wave number $k \to k_n$ to distinguish the solutions from each other. The $n = 0$ solution leads to the trivial solution of $\psi(x) \equiv 0$, so we discard that case. Plugging back in for B, we get for our n^{th} solution (which we also subscript with an n):

$$\psi_n(x) = A\left(e^{+ik_n x} - e^{-ik_n x}\right) = 2iA\sin\left(\frac{n\pi x}{L}\right), \quad n = 1,\, 2, \ldots \tag{6.1.6}$$

Well this certainly looks familiar! As we predicted, harmonic (plane wave) solutions inside the well, coupled with the requirement of nodes (vanishing probability) at the endpoints leads to a standing wave solution, with the harmonics determined by the n-values. We will see later that the interpretation of this "standing wave" is quite different from that of a standing wave on a string, but the math certainly matches.

Normalization

The reader may be troubled by the appearance of the "i" in the amplitude of our solution above. But there is no reason why A can't be complex as well. Keep in mind that all wave functions are equivalent up to a factor of a complex number with magnitude of 1, since all such wave functions give the same probability density. In any case, we *can* do better than just leaving the solution in this form, by using the normalization condition. Given that the wave function vanishes outside the well, the integral that usually goes from $x = -\infty$ to $x = +\infty$ can be reduced to an integral from 0 to L:

$$1 = \int_0^L \psi_n^*(x)\,\psi_n(x)\,dx = \int_0^L \left[-2iA^*\sin\left(\frac{n\pi x}{L}\right)\right]\left[2iA\sin\left(\frac{n\pi x}{L}\right)\right]dx = 4|A|^2\int_0^L \sin^2\left(\frac{n\pi x}{L}\right)dx \tag{6.1.7}$$

Performing the integral and solving for $|A|^2$ gives:

$$|A|^2 = \frac{1}{2L} \tag{6.1.8}$$

As we have said, the value of A is free to be anything that has this magnitude-squared, but it is traditional to choose the value of A that gives a real-valued amplitude for the standing wave, so choosing $A = \frac{-i}{\sqrt{2}}$ we get for the wave function that is the n_{th} harmonic:

$$\psi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right), \quad n = 1,\ 2, \ldots \tag{6.1.9}$$

Energy Spectrum

Now that we have the wave function for the stationary states, we can look into what measurable physical values we can expect to see. Highest on this list of observables is energy. Recall that the stationary-state solution gives us all of the eigenstates of energy, and the measurable values of energy are the eigenvalues associated with these states. We can therefore plug the wave function back into the Schrödinger equation for stationary states and solve for the possible values of E (the constant that appears in this equation:

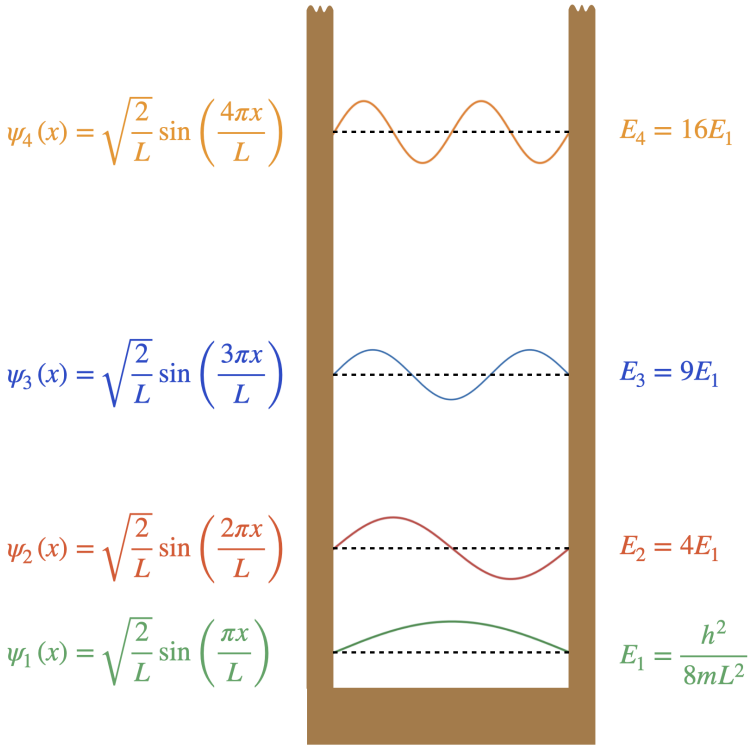
$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\left[\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\right] = +\frac{\hbar^2 n^2}{2\pi^2 m L^2}\left[\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\right] \tag{6.1.10}$$

The two derivatives on the sine function changed its sign, and brought out two factors of $\frac{n\pi x}{L}$. We can peel-off the constant in front of the wave function on the right-hand side of the equation, and set it equal to the energy. We see that the energy depends upon the harmonic n:

$$E_n = \frac{\hbar^2 n^2}{2\pi^2 m L^2} = \frac{h^2 n^2}{8mL^2} \tag{6.1.11}$$

We will abandon the use of the word "harmonic" in favor of *energy level*. The lowest energy level is referred-to as the *ground state*, and the energy levels above that are called *excited states*. So $n = 1$ corresponds to the ground state, $n = 2$ the first excited state, and so on.

Figure 6.1.2 – Energy Eigenstates and Eigenvalues of the Infinite Square Well



Physical Interpretation

Nothing seems particularly unusual about this solution until we think about how the result differs from what we expect to see classically. The first thing that comes to mind is that a we can start a ball bouncing elastically between two walls at any speed we wish, and therefore it can have any kinetic energy whatsoever. Certainly we would not expect to be able to only be able to measure certain allowable kinetic energies. While it is not yet clear why, it will turn out that this *quantization* of the energy spectrum is a general feature of all particles in bound states.

Another thing we find about the energy (and another thing that is true for bound states in general) is that the minimum energy level for the particle can never be the minimum potential energy of the well (i.e. it can never be found at the "bottom" of the well). One might be tempted to claim that the ground state can never have a zero value, but this is actually a silly statement, since we can set the zero-point of energy wherever we like. If we redefine our energy scale as $E' = E_n - E_1$, then the energy is zero at the ground state. But with this scale, the minimum potential energy is negative, so the ground state still doesn't get that low.

Next we consider momentum. For a ball bouncing back-and-forth elastically, we would expect to find it moving in either direction with equal probability, and with a fixed magnitude of momentum. We also see this in the quantum-mechanical case. However, considering what we found for kinetic energy, it's clear that we can't prepare the system with whatever fixed magnitude of momentum that we wish. Put another way, we will only measure certain magnitudes of momentum for the particle – half the time moving left and half the time moving right with a momentum magnitude of $|p_n| = \frac{hn}{2L}$.

In case you are wondering why it seems like we *can* start a ball bouncing back-and-forth between two walls with any momentum/KE we want, consider the tolerances we would need to measure to in order to prove it. A ball with a mass of $0.1kg$ can change its momentum in increments of $\frac{h}{2L}$, so if it is bouncing between walls separated by $20cm$, its "jump" in speed from one level to the next is:

$$v = \frac{p}{m} = \frac{h}{2mL} = \frac{6.63 \times 10^{-34} Js}{2\left(0.1kg\right)\left(0.2m\right)} = 1.66 \times 10^{-32} \frac{m}{s} \tag{6.1.12}$$

With such small increments of quantized speeds, it's no wonder it seems to us in the classical world that we can make the speed anything we want.

Possibly the strangest comparison between the classical and quantum results is the particle's position. Randomly measuring the position of a ball bouncing back-and-forth results in an uniform probability distribution. The ball moves at a constant speed, so naturally it spends the same amount of time in every small region Δx between the walls, making all these regions equally likely to find the ball. But a quantum particle in the ground state has a higher probability density at the center of the well than anywhere else, which means it is more likely to be found in a small region Δx near the center than in an equal-sized region closer to the walls. Stranger still, this behavior changes *dramatically* when the energy state is instead the first excited state. In this case, there is a node in the wave function at the center of the well, which means that unlike the ground state, for which the probability density is a maximum there, the probability density is actually *zero*.

6.2: Particle-in-a-Box, Part 2

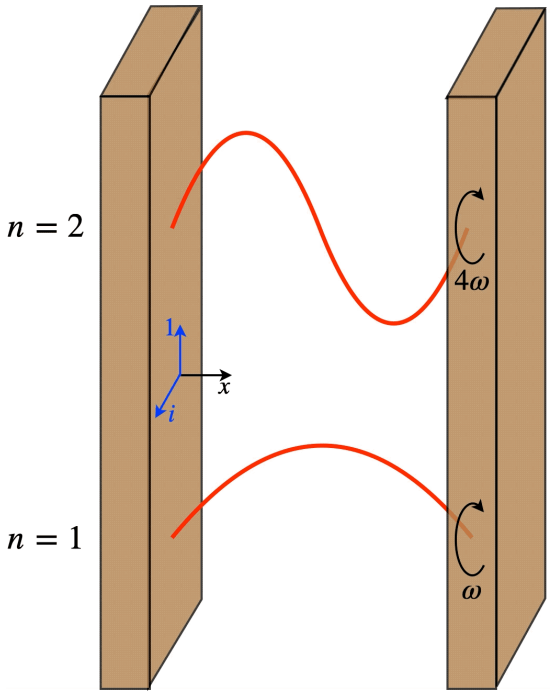
The Time-Dependent Stationary-State Wave Functions

It's tempting to think that everything about the infinite square well was solved in the previous section, but that was only the stationary states! Remember that there are infinitely-more solutions that can be built from those. A good starting point is to have a visual model for what is going on with these energy eigenstates. Even though we solved for the time-independent wave function $\psi(x)$, the *full* wave function for he quantum state of one of these energy eigenstates still has a time component to it:

$$\Psi_n(x,t) = \psi_n(x)\, e^{-i\omega_n t} = \sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right) e^{-i\frac{E_n}{\hbar}t}\,, \quad E_n = \frac{n^2h^2}{8mL^2} \tag{6.2.1}$$

We said that this is very similar to the solution of a standing wave on a string, and here we can see that the only difference is that it has both a real and imaginary sinusoid part to the time portion, while a standing wave on a string has only a real sinusoid for the time portion. So rather than having a mental picture of a string vibrating up-and-down, a better one is to think of a rotating jump-rope, with the antinodes swinging through the real and imaginary axes:

Figure 6.2.1 – Two Energy Eigenstates in a Box



These whirling jump-ropes represent the ground state and first-excited states of a particle in the infinite square well. There is one antinode for the ground state, and two for the first excited state, but the "oscillations" are not like those of a string standing wave. The figure defines "out of the page" as the positive imaginary axis, and "up" as the positive real axis (left-right is still the position (x) axis, with the left surface being $x = 0$ and the right $x = L$). Making this definition of the complex plane allows us to express the the time-dependent quantum phase as a rotation:

$$e^{-i\omega t} = \cos\omega t - i\sin\omega t \tag{6.2.2}$$

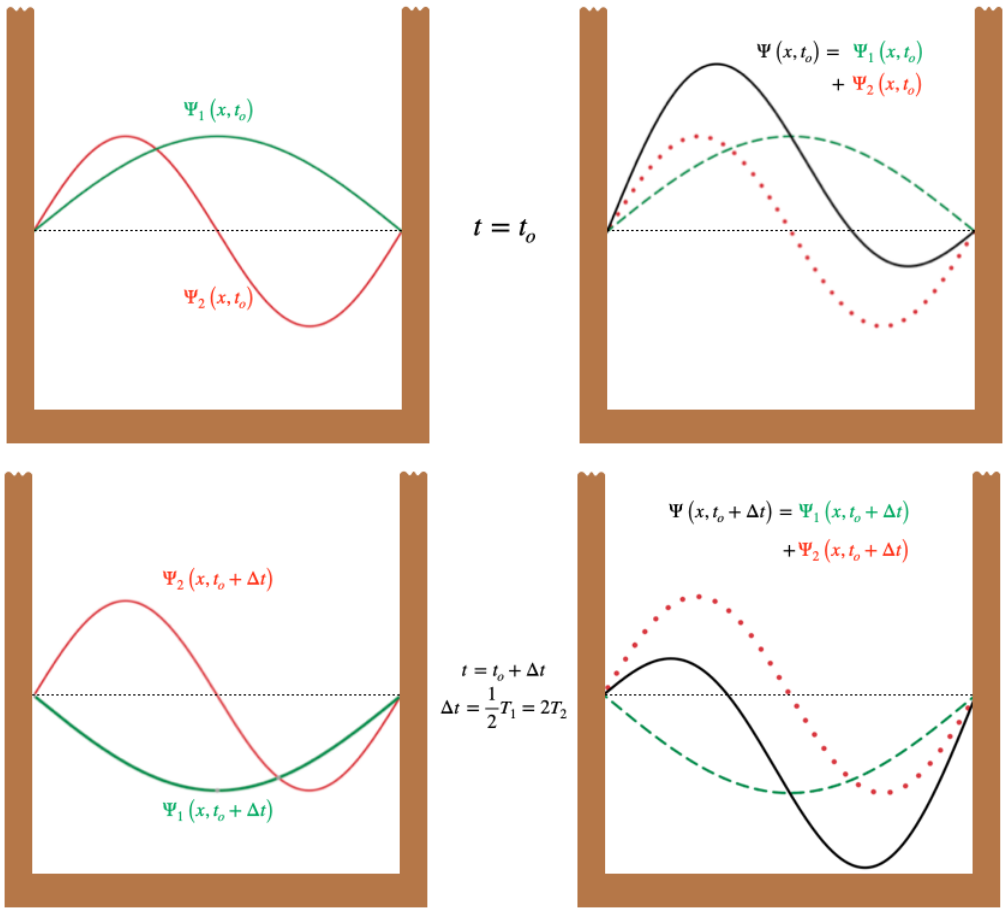
Notice that since $\omega_n \propto E_n \propto n^2$, the $n = 2$ state has 4 times the rotational velocity as the $n = 1$ state.

Given the obvious time-dependence of these solutions, the question arise of why they are called *stationary* states. It's important to keep in mind that it is only the probability *density* that comes into the calculation our measurements, not the probability amplitude. The magnitude of the wave function is represented in the figure above by the distance that the "jump-rope" is from the straight line joining the endpoints. Because the jump-rope is rotating, any given point (measured by x) on it is always the same distance from the center line, which means that the probability amplitude magnitude remains fixed in time at any given position x. With the probabilities of locating the particle at any position not changing, the state is indeed "stationary".

Non-Stationary-State Solutions

One might ask, "If the individual energy eigenstates are stationary, then why isn't a linear combination of them also stationary?" The answer lies once again in the figure above. While each pieces of a single jump rope maintains a constant distance between it and the center line, if we add (like vectors) the position of two jump ropes representing different energies (attached at the same endpoints), then the fact they rotate at different speeds will mean that the net displacement changes with time. For example, the $n = 2$ eigenstate rotates at 4 times the rate of the $n = 1$ state, so one-half period for Ψ_1 corresponds to two full periods for Ψ_2. This means that if the antinode of Ψ_1 is aligned in the same direction as the left antinode of Ψ_2 at some initial time, then one-half period for Ψ_1 later, its antinode will have flipped, while Ψ_2 returns to its initial state since two of its full periods elapse.

Figure 6.2.2 – Superposed Eigenstates at Two Times



When the two wave functions superpose at these two times, the amplitude is larger in the left side of the box initially, and then larger in the right side of the box later. When this amplitude is squared to get the probability density, we find that the particle is more likely to be found in the left half of the box at the initial time, and more likely to be found in the right half of the box afterward – certainly the probability is not "stationary" in this case!

We will now invoke the spectral theorem to write the general solution of the Schrödinger equation. This consists of a linear combination of the energy eigenstates, with coefficients that define the "recipe" of that general state:

$$\Psi(x,t) = \sum_{n=1}^{\infty} C_n \psi_n(x)\, e^{-i\omega_n t} = \sqrt{\frac{2}{L}} \sum_{n=1}^{\infty} C_n \sin\left(\frac{n\pi x}{L}\right) e^{-i\frac{E_n}{\hbar}t} \tag{6.2.3}$$

If we look at the wave form of Ψ at $t = 0$, we see a very familiar sum: It is a Fourier series! We know exactly how to calculate the coefficients in this case – we use the orthogonality property of the eigenstates, namely:

$$C_n = \int_0^L \psi_n^*(x)\, \Psi(x,0)\, dx \tag{6.2.4}$$

If the full wave function is known at some moment in time, then we can compute the entire eigenstate recipe (all of the C_n's) of that state. Once all of these are known, the expectation value of the energy for Ψ can be computed using Equation 5.4.8. Note that the resulting state must still have the properties of a wave function that satisfies the Schrödinger equation for this potential (i.e. it must vanish at the walls), and it must be normalized. The first of these requirements is automatically satisfied, since every term in the sum has this property, and the second is a restriction on the coefficients, as we have already noted in Equation 5.4.7.

Emission and Absorption

Ultimately we hope that models like the infinite square well will help us to explain the world around us. It's pretty rare that one-dimensional models will give us directly usable results, but they are nevertheless useful for gaining insight into what comes later. When it comes to the "real world," experiments need to be performed to determine (or confirm) energy spectra like the one we found for this boxed particle. How exactly do we do this? Well, what we observe is virtually always light. We trust the principle of conservation of energy, and a particle that is trapped in a box can presumably absorb or emit light if the change in its energy level is equal to the energy of the photon it absorbs or emits (obviously an absorption results in an increase of energy level, and an emission a decrease). The frequency of the photon is then:

$$hf = \Delta E = E_{n_2} - E_{n_1} \quad \Rightarrow \quad f = \frac{1}{h}\left(\frac{h^2 n_2^2}{8mL^2} - \frac{h^2 n_1^2}{8mL^2}\right) = \frac{h}{8mL^2}\left(n_2^2 - n_1^2\right) \tag{6.2.5}$$

The fact that the energy spectrum of the particle is quantized (the n's are integers) means that the spectrum of the light emitted by the particle comes only in discrete frequencies. Every model will have its own particular spectrum. This particular model has energy levels proportional to n^2, but we will see others during our journey, and the rough features of this spectrum are predictable based on certain qualities of the potential energy function.

6.3: The Finite Square Well

Lowering the Walls

As instructive as the infinite square well is, it's not particularly physical that its depth is infinite. Well okay, it works well as an approximation when the depth is much greater than the ground-state energy (so that lots of energy levels are available), but now we are going to look at a case when the particle is only loosely-held by a square potential. The finite square well we are about to discuss is a bit tougher to compare to a classical system (like the ball bouncing between two walls for the infinite case). In the case of the infinite square well, we could sloppily (and incorrectly) state that the particle remains confined to the well thanks to the infinite force at the walls (where the potential function has infinite slope). But the fact that the force is infinite simply means that the interaction time with the wall is infinitesimal, since there is a finite change in momentum for the particle. If that case, if the particle is moving faster, then the infinite force must be greater to deliver a greater impulse. The infinite depth of the well simply assures that there is always a larger impulse available, if needed, to turn the particle around, no matter how fast it is moving.

For a square well with a finite depth, the walls will still deliver infinite forces over infinitesimal time intervals to provide an impulse to the particle, but if the particle's kinetic energy exceeds the height of the top of the well, then the impulse by the wall will only serve to slow down the particle, and it will continue in the same direction, not restricted to the confines of the well. In this case, we simply have an unbound particle that speeds up when it is in the x range defined by the well, and slows down when it exits.

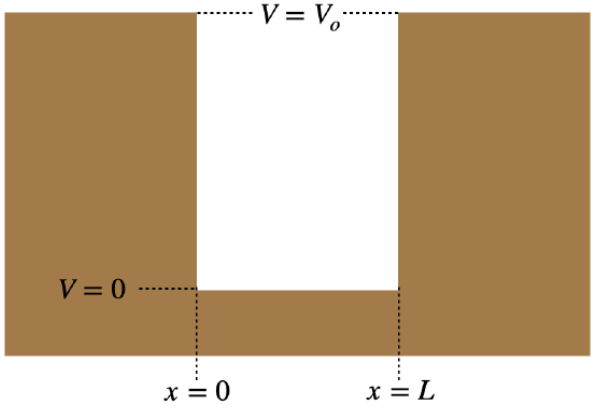
Having said all this, we are interested in the bound states (there is plenty of interesting quantum physics in the unbound case as well, but we will not cover that here). Keep in mind that the ground state is always above the bottom of the well, so if this well is particularly shallow, perhaps we will find that *no* bound states occur. For example, if we just look at the energy spectrum for the infinite square well, the ground state energy is $E_1 = \frac{h^2}{8mL^2}$, so it might be that if the well is shallower than this number (defined by the mass of the particle and width of the well), then there is no bound state at all. We won't know the answer to this until we get into the details, because we can't expect the energy spectrum of the finite square well to be identical to that of its bigger sibling.

Start with Schrödinger's Equation

We begin, as usual, looking for the stationary-state solutions using the separated Schrödinger equation for our new potential. The mathematical description of the potential looks very much like it did for the infinite square well, with the exception that the height is no longer infinite:

$$V(x) = \begin{cases} 0 & 0 \le x \le L \\ V_o & x < 0,\ x > L \end{cases} \tag{6.3.1}$$

Figure 6.3.1 – Finite Square Well



Plugging this potential into the time-independent Schrödinger equation, we create essentially two separate differential equations, as we did in the infinite square well case:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + \left\{ \begin{matrix} 0 & \text{(inside well)} \\ V_o & \text{(outside well)} \end{matrix} \right\} \psi(x) = E\psi(x) \tag{6.3.2}$$

As we said above, we are interested in the bound states, so we insist that $E < V_o$. We see at the outset that there will be a difference here from the infinite well. We started the infinite well by noting that the wave function must necessarily vanish outside the confines of the box, but it is not obvious that that is the case here. We know that classically the object can't ever be outside the confines of the well (its total energy is less than the potential energy, which makes the kinetic energy impossibly negative), but the mathematics does not rule out a non-zero probability amplitude immediately, and we've learned not to trust our classical reasoning by now. So let's skip worrying about the wave function beyond the walls for now, and continue with the process we followed with the infinite potential well – by writing down a general wave function for the "free" particle inside the well. Following the lead from before, we'll go with two plane wave states that are separable solutions to the free particle:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) = E\psi(x) \quad \Rightarrow \quad \psi_{\text{inside well}}(x) = Ae^{+ikx} + Be^{-ikx}\ , \quad E = \frac{\hbar^2 k^2}{2m} \tag{6.3.3}$$

As with the infinite well, these plane waves in the $V = 0$ region (for the solution to be steady-state) must be equally right- and left-moving, creating standing waves (these are the stationary-state solutions, after all!). We therefore take the "shortcut" of writing the internal wave function as a combination of sine and cosine functions:

$$\psi_{\text{inside well}}(x) = A\sin kx + B\cos kx\ , \quad E = \frac{\hbar^2 k^2}{2m} \tag{6.3.4}$$

Why does the cosine function make an appearance here, when it didn't show up for the infinite well, you ask? Well, the boundary conditions of the wave function vanishing at the boundaries is what led us to exclusive use sine in the previous case. Now with the boundary conditions allowing for a non-vanishing wave function, we need to account for both of these possibilities.

At this point in infinite square well case, we went immediately to the boundary conditions, but without the simplification of a vanishing wave function (nodes) at the endpoints, we need to exercise some restraint. That is, we cannot immediately relate the possible values of k to the length of the well L. So let's set this aside for now and have a closer look at the differential equation outside the walls:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) = (E - V_o)\psi(x) \tag{6.3.5}$$

Well, this is basically the same differential equation as inside the well, with E replaced with $E - V_o$. So can't we just use the same solution as above? Yes and no. The solution is still a sum of exponentials, but above the value of E is positive, while the value of $E - V_o$ is negative (the particle's total energy has to be less than V_o for it to be bound). If the wave function is a sum of sinusoids (like $e^{ix} = \cos x + i\sin x$), then two derivatives changes the sine of the wave function, and the differential equation works. But with a negative value multiplying $\psi(x)$ on the right-hand side of the equation, the result is that our exponentials lose their i's. (The reader is encouraged to confirm for themself that this wave function satisfies the differential equation):

$$\psi_{\text{outside well}}(x) = Ce^{+\alpha x} + De^{-\alpha x}\ , \quad V_o - E = \frac{\hbar^2\alpha^2}{2m} \tag{6.3.6}$$

Let's separate the outside-the-well regions into "left" and "right". The equation above applies to both, but we can reduce the terms required by invoking the fact that the wave function must ultimately be normalized. Looking at the wave function in the left region, we see that with $x < 0$, the term with a negative exponent will grow without bound as $x \to -\infty$. This requires that the coefficient D is identically zero. Similarly, the wave function in the right region will grow without bound as $x \to +\infty$ unless the constant C is identically zero.

Summarizing what we have so far for the wave function:

$$\psi(x) = \begin{cases} Ce^{+\alpha x} & x < 0 \\ A\sin kx + B\cos kx & 0 \le x \le L \\ De^{-\alpha x} & x > L \end{cases} \tag{6.3.7}$$

Energy Spectrum Quantization

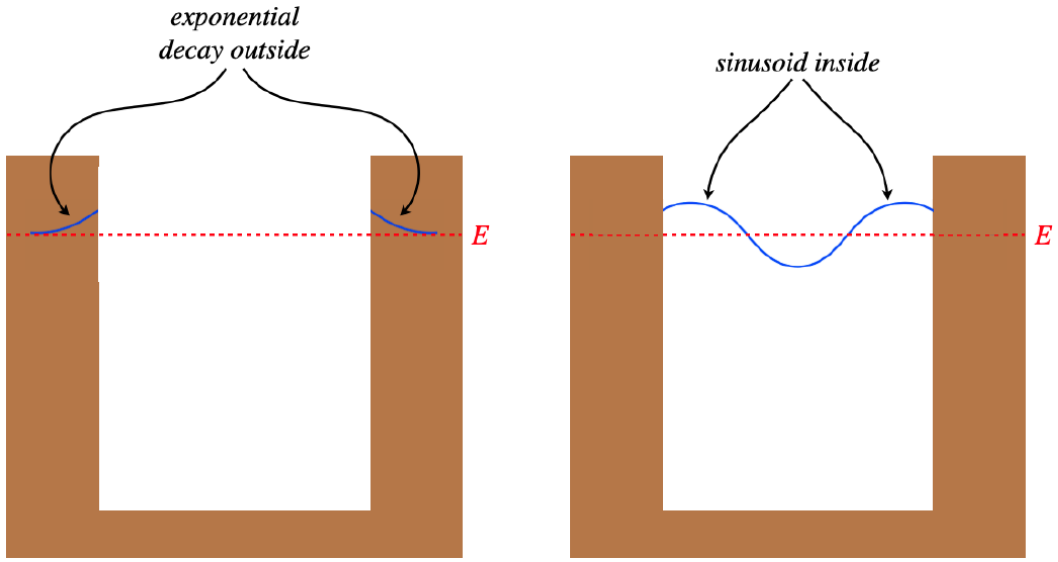
With the presence of a non-zero wave function outside the walls, our task of matching boundary conditions is more daunting than it was for the infinite well. We have two conditions that must hold at the boundaries. The first is that the wave function must be continuous – it is unphysical for the particle to have a sudden jump in its probability density. That is, an infinitesimal change in position needs to be accompanied by an infinitesimal change in probability of finding the particle. And the second requirement is

that the wave function's first derivative must also be continuous. If it is not, then the second derivative is not defined at the boundaries, throwing Schrödinger's equation into the garbage pail. This wasn't a problem when the infinite potential function covered-up this ugliness (the wave function was *identically* zero, which obviated any concern for its second derivative), but now we need to bow to the requirements of calculus.

So our task has now become setting the values and first derivatives of the wave functions equal on both sides of the borders $x = 0$ and $x = L$, and using these to find the unknown constants. Ultimately we are looking for the energy spectrum: E. This means that we need either k or α, and we need to find the in terms of the known values m (mass of the particle), L (length of the well), and V_o (depth of the well). Before we delve into this math, it is interesting to see conceptually why the energy spectrum in this case must be quantized, as we have previously claimed without proof to be a general feature of bound states.

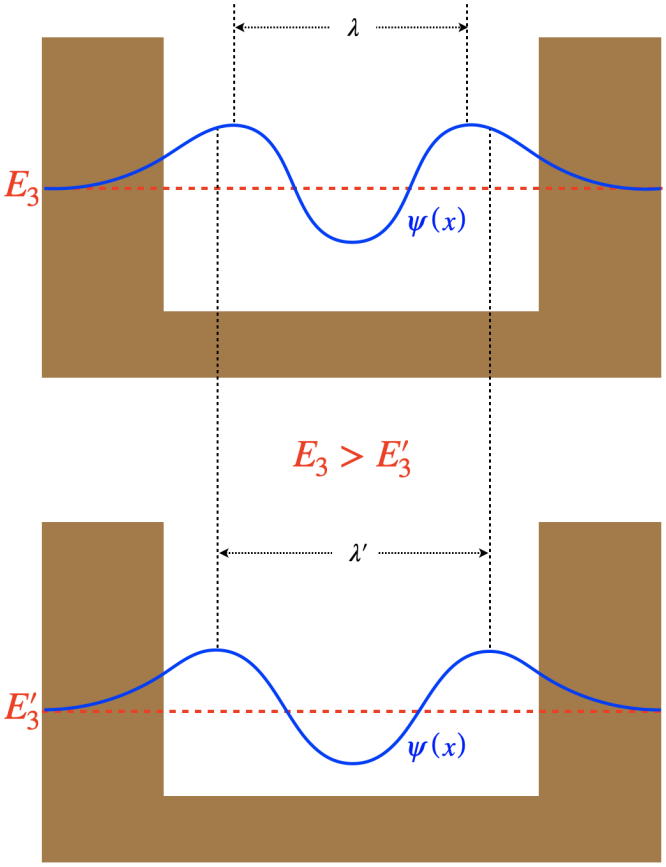
Consider an example of a wave function with three antinodes. What our math so far requires is:

Figure 6.3.2 – Requirements of Wave Function



This doesn't seem particularly restrictive. For example, we can conceive of two wave functions with different interior sinusoidal wavelengths (and therefore different wave numbers and energies) that satisfies these restrictions:

Figure 6.3.3 – Two Energies for n=3?



But there is a subtle flaw in these diagrams – looking closely reveals these waves don't remain "sinusoidal" all the way to the endpoints as they should (there should be no inflection point until the sinusoid reaches the axis). Let's see what it takes to satisfy this requirement. Let imagine constructing a ground-state wave function that satisfies the boundary conditions with an arbitrarily-chosen wavelength. We might proceed as follows:

1. Select an exponentially-decaying curve at the boundaries. Don't worry about the details yet, such as the value at the boundary or the decay constant.
2. Select a sinusoidal wave with our arbitrary wavelength. Well, it isn't totally arbitrary – to get a ground state, we need half this wavelength to be longer than the separation of the well walls, but this gives us a wavelength range of $L < \lambda < 2L$ to work with for the ground state.
3. Set the sinusoid into place, lowering it until the height of the sinusoid matches the height of the exponential at the boundary, satisfying the continuity of $\psi(x)$.
4. In general this will not allow the slopes of the exponential and sinusoid to match, so we can fix that by changing the amplitude of the sinusoid until the slopes do match.
5. Voilà! The wave function's boundary conditions match, and we selected the wavelength we wanted.

Figure 6.3.4 – A Scheme for Matching Boundary Conditions for a Given Wavelength

The last step, which seems like a trivial one, is to check to see if the wave function is fully normalized over the range $(-\infty, +\infty)$. Suppose the normalization integral comes out to be too small. Then we "just" have to raise the whole wave function (sinusoid and exponentials) together, making sure that we raise the points at the boundaries the same amount. But the problem is that these are two very different functions – changing parameters to increase the value at a specific point will not result in changing the slopes the same amount as well. If, for the moment, we call the center of the well the origin, then the sinusoid for the ground state is clearly just a cosine, and calling position of the right wall x_o, continuity of the wave function requires:

$$\psi(x_o) = A\cos kx_o = Be^{-\alpha x_o} \tag{6.3.8}$$

We can adjust A and B however we like to adjust the normalization, as long as this relation holds. Let's suppose that the changes we had to make to A and B to assure normalization were ΔA and ΔB, respectively:

$$\psi_{new}(x_o) = (A+\Delta A)\cos kx_o = (B+\Delta B)e^{-\alpha x_o} \quad \Rightarrow \quad \Delta A\cos kx_o = \Delta Be^{-\alpha x_o} \tag{6.3.9}$$

The original wave function was constructed to match the derivatives at the boundaries:

$$\psi'(x_o) = -Ak\sin kx_o = -B\alpha e^{-\alpha x_o} \tag{6.3.10}$$

For the derivatives to also match at the boundaries after the change of the values of A and B requires:

$$\psi'(x_o) = -(A+\Delta A)k\sin kx_o = -(B+\Delta B)\alpha e^{-\alpha x_o} \quad \Rightarrow \quad -k\Delta A\sin kx_o = -\alpha\Delta Be^{-\alpha x_o} \tag{6.3.11}$$

Dividing the two equations we obtained reveals that both boundary conditions remain matched after the shifts of ΔA and ΔB only under specific conditions related to k and α:

$$\frac{-\Delta Ak\sin kx_o}{\Delta A\cos kx_o} = \frac{-\alpha\Delta B\alpha e^{-\alpha x_o}}{\Delta Be^{-\alpha x_o}} \quad \Rightarrow \quad k\tan kx_o = \alpha \tag{6.3.12}$$

In short, while we could make the boundary conditions match with this scheme, we can't also assure the most essential feature of the wave function – that it be normalized – unless the wave has a very specific wavelength. This means that like the infinite square well, the energy spectrum of the finite square well is quantized.

The Math

All that remains to this problem is to apply all the boundary conditions to obtain the energy spectrum and energy eigenfunctions. Well, in principle this is the idea, but unlike the infinite square well, where the energy eigenvalues are a simple function of the eigenstate number n, the math does not behave so nicely here, as we will see...

There are four boundary conditions here – two boundaries, and two conditions each:

$\psi(x)$ continuous at $x = 0$

$\psi(0) = Ce^0 = A\sin 0 + B\cos 0 \quad \Rightarrow \quad C = B$

$\psi(x)$ continuous at $x = L$

$\psi(L) = De^{-\alpha L} = A\sin kL + B\cos kL$

$\frac{d}{dx}\psi(x)$ continuous at $x = 0$

$$\psi'(0) = \alpha C e^0 = kA\cos 0 - kB\sin 0 \quad \Rightarrow \quad \alpha C = kA$$

$\frac{d}{dx}\psi(x)$ continuous at $x = L$

$$\psi'(L) = -\alpha D e^{-\alpha L} = kA\cos kL - kB\sin kL$$

We have four equations and four constants to eliminate. Leaving the tedious algebra as an exercise for the reader, we end up with:

$$\tan kL = \frac{2\alpha k}{k^2 - \alpha^2}\,, \quad k = \frac{\sqrt{2mE}}{\hbar}\,, \quad \alpha = \frac{\sqrt{2m\left(V_o - E\right)}}{\hbar} \tag{6.3.13}$$

One would of course like to solve these for the energy E in terms of m, L, and V_o, but we cannot do this in closed-form for this transcendental equation. The periodic nature of the tangent function ensures that the spectrum is quantized, but numerical/graphical methods show that there is a limited number of these solutions, depending upon the energy of the particle and the depth of the well.

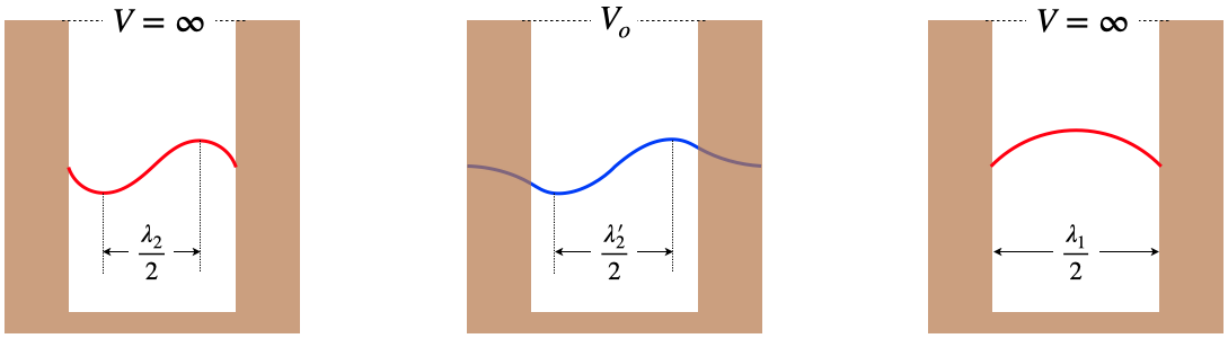
Effects of Varying the Well Depth

It is instructive to summarize what happens to the stationary state solutions and energy eigenvalues for a given particle in a well of a fixed length, as the depth of the well grows. The first thing that we note the states are characterized by the number of antinodes between the walls. In the case of the infinitely-high walls, there were nodes at the endpoints, but even though that restriction is lifted for the finite well, we can use this same criterion to determine if an eigenstate exists for the well, as follows:

Suppose there are 2 antinodes between the walls. This puts limits on the wavelength that the sinusoid can have. The wavelength must be no shorter than L, and no longer than $2L$. As we are holding the particle mass and well length fixed, one quantity that comes up frequently in these calculations is the ground state energy of the same particle in an infinite well of the same length, so we will scale all the energies of the finite well using this constant:

$$\epsilon \equiv \frac{h^2}{8mL^2} \tag{6.3.14}$$

Figure 6.3.5 – Limits of Wavelength for 2 Antinode Wave Function

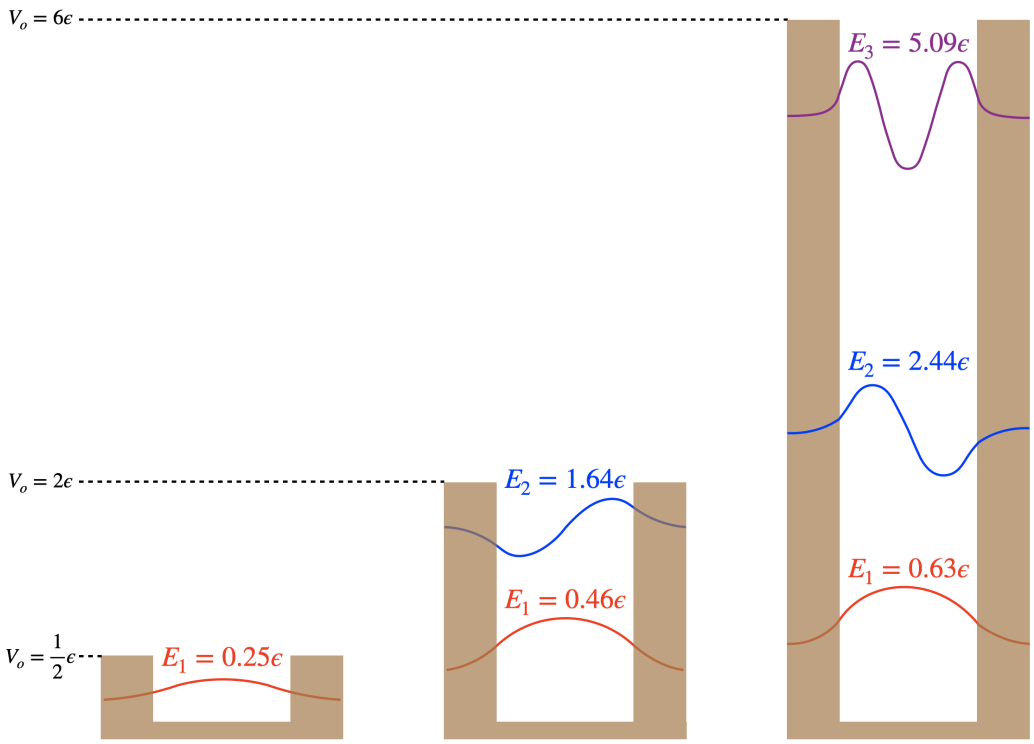


$$E_2 = \frac{h^2}{2m\lambda_2^2} = \frac{h^2}{2mL^2} = 4\epsilon \qquad 4\epsilon > E_2' = \frac{h^2}{2m\lambda_2'^2} > 1\epsilon \qquad E_1 = \frac{h^2}{2m\lambda_1^2} = \frac{h^2}{8mL^2} = 1\epsilon$$

This clearly limits the number of eigenstates for a given value of V_o. In particular, there can exist no bound eigenstate with an energy greater than V_o, and the eigenstate energy is determined by the wavelength. So the highest possible energy eigenstate for the finite well is the one with the lowest n for which $V_o < n^2\epsilon$.

The figure below shows solutions generated through numerical means, and depicts how the energy spectrum of a finite square well changes when only its depth is changed (while the particle mass and length of the well remains fixed).

Figure 6.3.6 – A Few Well Depths



Some things to note:

- The energy of the n^{th} eigenstate rises as V_o increases, and in the limit of $V_o \to \infty$ converges to $E_n = n^2\epsilon$.
- As V_o increases, a greater fraction of the wave function for a given eigenstate exists within the well (i.e. the probability of finding the particle in the classically-forbidden region drops). In the limit of $V_o \to \infty$, this probability drops to zero.
- As noted above, the energy of the n^{th} eigenstate always falls between $(n-1)^2\epsilon$ and $n^2\epsilon$. As the well is made deeper, it "picks up" new eigenstates when V_o crosses those special $n^2\epsilon$ values.

The Classically-Forbidden Region

A few more words need to be said about the non-zero probability of finding the particle in the region that is forbidden by classical physics. The constant in the exponential decaying probability is, from Equation 6.3.6:

$$\alpha = \frac{\sqrt{2m\,(V_o - E)}}{\hbar} \tag{6.3.15}$$

We can see immediately that when the walls of the potential well are infinitely-high, this value goes to ∞, and this has the effect of decaying the wave function to zero immediately, leaving no wave function outside the well, as we expect:

$$\lim_{V_o \to \infty} Ae^{-\alpha|x|} = 0 \tag{6.3.16}$$

But for finite values of V_o, the particle will not be so constrained, and it should be clear that α makes for a good proxy for how far the wave function "penetrates" into the classically forbidden region. The smaller α is (i.e. the closer E is to V_o) is, the more probable it is that the particle will be found outside the well. It is conventional to express this property as a characteristic distance, called the *penetration depth*, equal to the inverse of the value of α (which has units of length^{-1}):

$$\delta \equiv \alpha^{-1} = \frac{\hbar}{\sqrt{2m\,(V_o - E)}} \tag{6.3.17}$$

We can relate the penetration depth to an expectation value – something that we can measure experimentally. Suppose we make many measurements of the position of a particle that is restarted in the same state, and then only keep the results where the particle was found outside the well. We can then average the distances from the wall, to get an expectation value of the position of the particle *given it has penetrated into this region*. We can see how this expectation relates to the penetration depth δ with a simple

calculation. For simplicity, we will define the right wall's position as $x = 0$ (the length of the well is still L and the particle's mass is still m). Then the unnormalized wave function for positive values of x is:

$$\psi(x) = Ae^{-\alpha x}, \quad x \geq 0 \tag{6.3.18}$$

Normalizing (so that probabilities and then expectation values work out properly) gives:

$$1 = \int_0^\infty |\psi(x)|^2 dx = \int_0^\infty |A|^2 e^{-2\alpha x} dx \quad \Rightarrow \quad A = \sqrt{2\alpha} \tag{6.3.19}$$

And now the expectation value of x in this $x \geq 0$ region is:

$$\langle x \rangle = \int_0^\infty x \left[\sqrt{2\alpha} e^{-\alpha x}\right]^2 dx = 2\alpha \int_0^\infty x e^{-2\alpha x} dx = \frac{1}{2\alpha} = \frac{1}{2}\delta \tag{6.3.20}$$

So on average, when we look only at locations of the particle in the classically-forbidden range, we find it at a position equal to half of what we have defined as the penetration depth. We can of course also calculate the uncertainty of this measurement. The details of the integration are not diffficult, and are left to the reader, but the result is:

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \frac{1}{2}\delta \tag{6.3.21}$$

So treating the uncertainty as a range over which we have some sense of confidence in our experiment, we find that given we measure the particle in the classically forbidden region, we are confident that it ended up somewhere in the region from the wall to the penetration depth.

Okay, so let's close by addressing the confusing question of how a particle could be found in this region at all, given it would have to have negative kinetic energy to be in that region. If we measure the momentum of the particle in one of the eigenstates inside the well, then of course all we ever see is the plane-wave momentum, which means it has a well-defined kinetic energy as well. But outside the well, the momentum is not well-defined, as the wave function is not sinusoidal, and is generally a (Fourier) mix of plane waves associated with many possible momenta. There is a finite uncertainty in the position of the particle in this region (we just computed it), so the uncertainty principle tells us that there is a minimum uncertainty in the momentum in that region as well:

$$\Delta p \geq \frac{\hbar}{2\Delta x} = \frac{\hbar}{\delta} \tag{6.3.22}$$

The kinetic energy of the particle that has this uncertain momentum is also uncertain. Without working out the math, we can say estimate the uncertainty in this kinetic energy. Let's just define the "range" of momentum as being between $\langle p \rangle - \Delta p$ and $\langle p \rangle + \Delta p$. We can compute the minimum and maximum values of the kinetic energy in this range. While this minimum and maximum do not define the precise uncertainty for the kinetic, it is a good number to use as a lower limit of this uncertainty. This means that the uncertainty of the kinetic energy is:

$$\Delta KE > \frac{\Delta p^2}{2m} = \frac{\hbar^2}{2m\delta^2} = V_o - E \tag{6.3.23}$$

So the uncertainty in the kinetic energy measurement is greater than what we calculate to be the entire negative kinetic energy (using the kinetic energy of the particle inside the well), meaning that our measurement does not confirm to within uncertainty that the kinetic energy had to become negative.

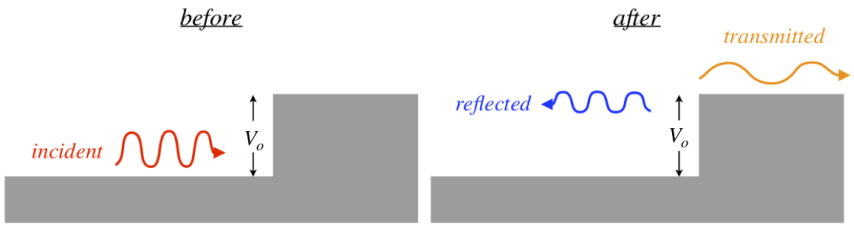
6.4: Tunneling

The Potential Step

In our examination of square wells, we noted that inside the well the wave function is a superposition of left- and right-moving plane waves. In these cases, these plane waves had equal wavelengths, because the energy was confined to a region. Here we will consider the possibility that when a plane wave strikes a vertical potential change, only *some* of the wave is reflected, while some of it is transmitted. Indeed, in our study of light, we found that this occurs. The upshot is that we cannot write our stationary-state solutions as sines and cosines as we did in the case where the left- and right-going plane waves had equal amplitudes.

We will remain with the stationary-state solution, meaning that the probabilities don't change over time, and the total energy remains fixed. The picture is a little more confusing here, as we will first consider a "before/after" scenario (which is hardly time-independent!), but eventually we will turn this into a steady-state situation, where "before" and "after" are occurring simultaneously and continuously. We will do this little-by-little, starting with the most basic cases and building our way up. The first case involves a totally-free particle that "encounters" a sudden change in the constant potential (a "step"), from 0 to V_o..When the wave encounters this step, it's reasonable from our understanding of the behavior of waves to assume that some of the wave will be reflected backward, while some is transmitted forward. A picture you might have in your head for this physical situation is:

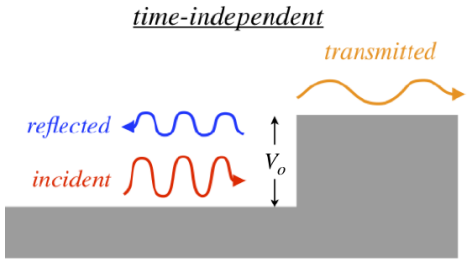
Figure 6.4.1 – Step Potential Before/After Picture



While this makes sense from a classical standpoint – moving objects "encounter" things in their travels all the time – it's a bit troublesome for quantum mechanics, as a plane wave occupies all values of x at once, so how can it be moving along the x direction? Put another way, we are seeking a *stationary state* solution, and this picture clearly shows a system evolving over time.

We therefore are looking for a solution where the wave function is unchanging over all values of x, but in practical terms we know that particles (which travel in wave packets of many energies rather than perfect plane waves) are localized, and therefore do really encounter changing potentials. What compromise can we reach in this regard? As stated above, any combination of plane waves going in both directions with the same momentum (for a given potential) will be an energy eigenstate, so how about if we restrict ourselves to only right-moving plane waves on the "transmitted" side of the step, and plane waves in both directions on the incident/reflected side? We can interpret this in our "real world" setting as a steady-state circumstance, where many particles are coming in, and some are reflected while others are transmitted, where the fractions of each are determined (mainly) by the relative amplitudes of the reflected and transmitted waves. The correct picture then becomes:

Figure 6.4.2 – Step Potential Steady-State Picture



Keep in mind that we are assuming the particle remains free, so even after the increase in potential energy, it has kinetic energy (albeit less). Clearly three is the minimum number of wave functions that are possible. For example, it is neither possible to have all of the wave to be reflected, nor have all of the wave to be transmitted, thanks to the finite value of V_o that is smaller than the particle's total energy. It *is* possible to get a solution that includes a left-moving wave on the right of the step, but we have discarded these solutions to fit with our narrative that there is a steady source of incoming particles from the left.

Okay, with all that out of the way, we follow our usual procedure of writing the wave function in each of the regions as superpositions of plane waves of appropriate energy, with unknown amplitudes. We'll define the position of the step as $x = 0$, which gives:

$$\psi(x) = \begin{cases} \psi_{inc}(x) + \psi_{refl}(x) & x < 0 \\ \psi_{trans}(x) & x > 0 \end{cases} \tag{6.4.1}$$

The incident and reflected plane waves experience the same potential, so they have the same wave number (k), while the wave number of the transmitted plane wave is different (k'). Also, they all have different amplitudes in general, so:

$$\left.\begin{array}{l}\psi_{inc}(x)=Ae^{+ikx}\\ \psi_{refl}(x)=Be^{-ikx}\end{array}\right\}\quad k=\frac{\sqrt{2mE}}{\hbar}$$
$$\psi_{trans}(x)=Ce^{+ik'x}\qquad k'=\frac{\sqrt{2m(E-V_o)}}{\hbar} \tag{6.4.2}$$

Let's stop a moment to interpret the quantities A, B and C. Suppose we were asked the probability density of finding an incident particle at a position between x and $x+dx$. The answer would be the magnitude-squared of its wave function. The same is true for a reflected or transmitted particle:

$$\text{probability of finding between } x \text{ and } x+dx: \begin{cases}\text{incident particle:} & |\psi_{inc}(x)|^2dx &=& |A|^2dx\\ \text{reflected particle:} & |\psi_{refl}(x)|^2dx &=& |B|^2dx\\ \text{transmitted particle:} & |\psi_{trans}(x)|^2dx &=& |C|^2dx\end{cases} \tag{6.4.3}$$

We would like to compare these quantities to determine the relative probabilities of the reflected and transmitted waves. In our steady-state model where many particles are coming in, this would tell us what fraction of the incoming particles are reflected, and what fraction is transmitted. The trouble is, these cannot be compared directly, because they have different wavelengths. Why is this a problem, and how do we resolve it?

Huygens's principle states that a wave propagates by having a lead crest generate more crests which propagate by generating more crests, and so on. It therefore stands to reason that one cycle of the incident incoming wave is responsible for one cycle of the reflected and one cycle of the transmitted wave. Suppose that the incident and reflected waves (which have the same wavelength) have a wavelength one half as long as the transmitted wave (which must be longer, as it has less kinetic energy and therefore less momentum). That means that two full cycles of the incident and reflected waves fit into the same space as the transmitted wave, so the particle is *twice as likely* to be found to the left of the barrier as in an equal space to the right. So to compare probabilities, we need to divide each of the probability densities by their associated wavelengths (or equivalently, multiply them by their associated wave numbers) in order to make a proper comparison. [In our steady-state many-particle model, this is the equivalent of accounting for the speed at which the particles are moving into and out of the step.]

We therefore define the *transmission and reflection probabilities* as the ratios of the relevant (adjusted) probability densities:

$$T=\frac{|C|^2k'}{|A|^2k}\qquad R=\frac{|B|^2k}{|A|^2k}=\frac{|B|^2}{|A|^2} \tag{6.4.4}$$

All we need now are A, B and C. We get these by matching the boundary conditions of the wave functions in the two regions – continuous (equal value) and smooth (equal derivative) – at $x=0$.

Sparing you the algebra, we get:

$$T=\frac{4\sqrt{E(E-V_o)}}{\left(\sqrt{E}+\sqrt{E-V_o}\right)^2}\qquad R=\frac{\left(\sqrt{E}-\sqrt{E-V_o}\right)^2}{\left(\sqrt{E}+\sqrt{E-V_o}\right)^2} \tag{6.4.5}$$

It is easy to show that the sum of these two probabilities comes out to one, which is consistent with the requirement that the incident particle must either reflect or be transmitted.

It should be pointed out that a similar solution results from a step *down*. Intuitively, it might seem like the wave only has a partial reflection when it steps up, because that seems like an obstacle, while stepping down is "easy." But even in a classical study of waves, one finds that waves reflect off any interface between media that result in different wave speeds, whether the speed changes to a slower speed or a faster one.

Speed Bumps and Potholes

We can extend this analysis to the case of a square potential bump (which isn't higher than the particle's energy) or dip (of any depth), where the potential before and after the obstruction is the same. Unfortunately, the stakes are raised, in that we now have *two* transitions – one at the front and one at the back of the obstruction. Following the reasoning above, we throw out the left-moving part of the wave function on the side opposite the right-moving incident wave. But we *cannot* throw out the reflected wave at the back surface of the obstruction. This leaves us with 5 parts of the wave function: The incident wave, the reflected wave, the two oppositely-moving waves in the region of the obstruction, and the transmitted wave.

This gives us 5 amplitudes to deal with, and two conditions at each boundary. Also, there is another parameter involved – the width of the obstruction. We are not interested in the specifics of what is going on in the region of the obstruction, only the transmission and reflection probabilities, which means there is one element of this problem that is simpler than the potential step: We don't have to account for different wavelengths of the incoming and transmitted waves, since we are assuming the starting and ending potential energies are the same.

Once again skipping the rather daunting amount of algebra, we get the following probabilities. Calling the width of the obstruction L and noting that the solutions involving a difference ($E-V_o$) are for potential bumps of height V_o, while those involving the sum ($E+V_o$) are for potential

dips of depth V_o:

$$T=\frac{4\left(\frac{E}{V_o}\right)\left(\frac{E}{V_o}\pm 1\right)}{\sin^2\left(\sqrt{2m\left(E\pm V_o\right)}\frac{L}{\hbar}\right)+4\left(\frac{E}{V_o}\right)\left(\frac{E}{V_o}\pm 1\right)} \qquad R=\frac{\sin^2\left(\sqrt{2m\left(E\pm V_o\right)}\frac{L}{\hbar}\right)}{\sin^2\left(\sqrt{2m\left(E\pm V_o\right)}\frac{L}{\hbar}\right)+4\left(\frac{E}{V_o}\right)\left(\frac{E}{V_o}\pm 1\right)} \tag{6.4.6}$$

Once again, the sum of these two probabilities is one, which means the particle doesn't have any chance of being trapped within the obstruction indefinitely. Or, put in terms of our steady stream of particles, none are left in the obstruction, so every particle that reaches the obstruction comes out in one direction or the other.

There is a fascinating special case that comes from this solution. It is possible for the numerator of the reflection probability to be zero, which means that *all the particles are transmitted*. This occurs when the argument of the sine function is a multiple of π:

$$\sqrt{2m\left(E\pm V_o\right)}\frac{L}{\hbar}=n\pi \quad \Rightarrow \quad E\pm V_o=\frac{n^2\pi^2\hbar^2}{2mL^2}=\frac{n^2h^2}{8mL^2} \tag{6.4.7}$$

Notice that $E\pm V_o$ is the kinetic energy of the particle within the obstruction – it increases in the case of a dip ($E+V_o$), and decreases in the case of a bump ($E-V_o$). It is a plane wave, so the kinetic energy can be written in terms of the wave number, which can then be written in terms of the wavelength, giving:

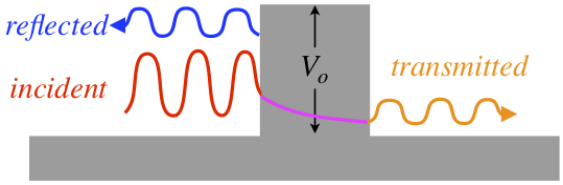
$$E\pm V_o=\frac{\hbar^2k'^2}{2m}=\frac{4\hbar^2\pi^2}{2m\lambda'^2}=\frac{n^2\pi^2\hbar^2}{2mL^2} \quad \Rightarrow \quad L=\frac{n\lambda'}{2} \tag{6.4.8}$$

Obstructions with thicknesses that are an integer number of half-wavelengths of the wave function measured within the obstruction are totally transparent to the beam of particles. We see this phenomenon with light, in the topic of thin films, which includes applications such as camera lens coatings.

Tunneling

We at last come to the quantum-mechanically iconic phenomenon of tunneling. In the case of a bump above, we assumed that the height of the bump was lower than than the energy of the particle. We now assume that the potential increase of the barrier, while finite, is greater than the energy of the incoming particle. As we already know, with a wall of finite height, some of the wave function "leaks" into the wall, exponentially decaying with respect to the penetration distance. Although it decays, it doesn't go to zero in the finite distance that is the thickness of the wall. Matching the boundary conditions on the other side of the wall results in a non-vanishing free particle wave function on the opposite side.

Figure 6.4.3 – Tunneling Through a Barrier



We have a remarkable shortcut to get us to the transmission and reflection probabilities. The math for this case is identical to the math used to derive T and R for the classically-surmountable bump above, with the exception that the kinetic energy $E-V_o$ within the barrier is negative. This appears within a square root, so we can introduce an imaginary number thus:

$$\sin^2\left(\sqrt{2m\left(E-V_o\right)}\frac{L}{\hbar}\right)=\sin^2\left(i\sqrt{2m\left(V_o-E\right)}\frac{L}{\hbar}\right) \tag{6.4.9}$$

Sine functions with imaginary arguments can be converted to hyperbolic sine functions multiplied by imaginary i. The proof is quick:

$$\sin x=\frac{e^{ix}-e^{-ix}}{2i} \quad \Rightarrow \quad \sin ix=\frac{e^{-x}-e^{x}}{2i}=i\frac{e^{x}-e^{-x}}{2}=i\sinh x \tag{6.4.10}$$

This converts the square of the sine function into the negative of the square of the sinh function, and since the sign of $\frac{E}{V_o}-1$ also flips when $\frac{E}{V_o}<1$, the negative sign that appears in both numerator and denominator cancel, leaving:

$$T=\frac{4\left(\frac{E}{V_o}\right)\left(1-\frac{E}{V_o}\right)}{\sinh^2\left(\sqrt{2m\left(V_o-E\right)}\frac{L}{\hbar}\right)+4\left(\frac{E}{V_o}\right)\left(1-\frac{E}{V_o}\right)} \qquad R=\frac{\sinh^2\left(\sqrt{2m\left(V_o-E\right)}\frac{L}{\hbar}\right)}{\sinh^2\left(\sqrt{2m\left(V_o-E\right)}\frac{L}{\hbar}\right)+4\left(\frac{E}{V_o}\right)\left(1-\frac{E}{V_o}\right)} \tag{6.4.11}$$

Unlike the case of a small potential bump, there are no special wavelengths that allow the particle to pass through without any reflecting. Not surprisingly, the transmission rate rises as the energy of the particle rises, and drops as the barrier's height or width increases.

6.5: The Quantum Harmonic Oscillator

Basic Features

As we did with the particle-in-a-box, we'll start with a review of the basic features of the quantum harmonic oscillator. Unlike the particle-in-a-box, the first treatment of this potential didn't include the position-space wave functions (other than their general features), so this review will be quite brief. Let's start with the stationary-state Schrödinger equation in position-space:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi_n(x)+\frac{1}{2}\kappa x^2\ \psi_n(x)=E_n\ \psi_n(x) \tag{6.5.1}$$

Alert

Note that the spring constant for this potential is represented by the greek letter kappa (κ), to distinguish it from the ubiquitous variable k that we use to represent the wave number.

We can employ many of the properties of wave functions and their energy spectra to get some sense of what the wave functions for the energy eigenstates look like:

1. **potential is infinite** – Like the infinite square well, this potential will have an infinite number of energy levels.
2. **energy levels will be quantized and ground state is non-zero** – Something we see for all bound states. As with the other wells we have seen, this comes about because we have to fit the interior wave function perfectly between the barriers while matching boundary conditions. This is why we introduced the "n" as a subscript to the wave function and the energy eigenvalues.
3. **parity flips every time we go up another energy level** – The ground state should be an even function, the first excited state an odd function, etc.
4. **potential grows to infinity, but for any given energy level the "wall" is finite** – The boundary conditions for wave function does not require that it vanish at the classical stopping points, as it did for the box, because the walls are not infinitely-high at the points where the classically-forbidden region begins. The wave function should therefore "leak" into the walls, giving the particle a non-zero probability of being found in the classically-forbidden region.

There are some ways that these wave functions should differ from those for the infinite square well:

1. **gap between the walls grows as the energy level grows** – As usual, an antinode is added every time we go up an energy level, in order to alternate between even and odd functions. For the infinite square well, this was easy to account for, since the distance between the walls never changed. The wavelength change in going from level n to level $n+1$ was a reduction by a factor of $\frac{n}{n+1}$. But for the harmonic oscillator potential, the classical turning points get farther apart as the energy grows. So while each energy level requires an additional half-wavelength, those wavelengths don't need to shrink as fast as the levels rise, in order to fit between the turning points. This means that the jumps between energy levels will not be as great for the harmonic oscillator as they were for the infinite square well (which was proportional to n^2).
2. **classical limit (very high energy levels) is different from infinite square well** – As the energy levels in the box get higher, the number of antinodes increase, and at very high energies, there are antinodes virtually everywhere within the box. Every antinode corresponds to an equal probability amplitude, so at very high energies the probability distribution is uniform. The high-energy limit is called the *classical limit*, and indeed for the box we get the correct result. We don't yet know what the wave functions will look like for the harmonic oscillator, but classically we do not expect the probability distribution to be uniform for a mass on a spring, as the mass spends significantly more time near the turning points than near the center. So the stationary-state wave functions for very high energies will not converge to a uniform distribution, and in fact should peak at the classical turning points.
3. **potential is not constant within the well** – For the infinite square well, within the well, the potential is constant (zero), making the solution in that region a combination of two opposite-moving plane waves. In this case, the potential changes continually, so we expect that we'll need to sum an infinite number of plane waves. It isn't clear what the spectral content of the full stationary state wave function will be, and we won't solve this problem from scratch, but we will examine the solution nonetheless, as it has some illuminating features.

Wave Functions

A solution of Equation 6.5.1 with proper boundary conditions yields stationary-state wave functions and an energy spectrum consistent with the above observations. Solving this differential equation "from scratch" gets too far into the weeds mathematically,

but we can make some educated guesses, and work our way "backwards" to the rest. As a start, we note that the we need two derivatives to give back the wave function itself, multiplied by a constant (the energy eigenvalue) plus a function of x^2 (the potential energy term). Whenever the wave function mus return, we think of an exponential. The problem is getting an x^2 factor from two derivatives. Note that if the wave function has an x^2 *in the exponent*, then a single derivative will bring down a$2x$ from the chain rule. Another such derivative will bring down a second factor of $2x$, giving us the x^2 we need, and the product rule resulting from the second derivative will also give a constant factor. So let's try this wave function:

$$\psi(x) = Ae^{-\alpha x^2} \tag{6.5.2}$$

This is just the general form that we are trying. To extract more information, we need to plug it into Schrödinger's equation and see what comes out. We will also need to eventually normalize it, so it can be used to compute probabilities, expectations values, etc. So putting this into Equation 6.5.1 gives:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\left[Ae^{-\alpha x^2}\right] + \frac{1}{2}\kappa x^2\left[Ae^{-\alpha x^2}\right] = E_n\left[Ae^{-\alpha x^2}\right] \tag{6.5.3}$$

Taking the derivatives and canceling the \(e^{-\alpha x^2}\) functions that appear in every term gives:

$$-\frac{\hbar^2}{2m}\left(-2\alpha + 2\alpha^2 x^2\right) + \frac{1}{2}\kappa x^2 = E_n \tag{6.5.4}$$

For $\psi(x)$ to be a solution to the Schrödinger equation, this equation has to hold for all values of x. This means that the coefficients of x^2 must cancel:

$$0 = -\frac{\hbar^2}{m}\alpha^2 + \frac{1}{2}\kappa \quad \Rightarrow \quad \alpha = \frac{\sqrt{\kappa m}}{2\hbar} \tag{6.5.5}$$

With the x^2 terms canceling, the constant terms are left behind, giving:

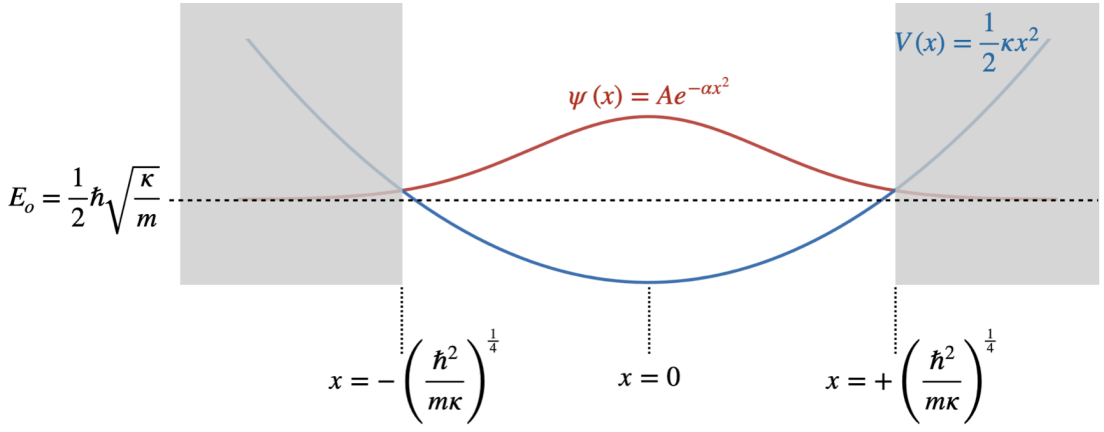
$$E_n = \frac{\hbar^2\alpha}{m} \tag{6.5.6}$$

Putting together these two results gives the energy in terms of given values:

$$E_n = \frac{\hbar^2\frac{\sqrt{\kappa m}}{2\hbar}}{m} = \frac{1}{2}\hbar\sqrt{\frac{\kappa}{m}} \tag{6.5.7}$$

Wait a second. Where is the n on the right side of the equation? The answer is that this choice of $\psi(x)$ solves the differential equation, so it is a wave function for an eigenstate of energy, but only one – unlike the particle-in-a-box, we are not able to use the "nodes must exist at both ends" criterion to get all of the eigenstates at once. Okay, so which eigenstate is this? We can answer this using the knowledge that the eigenstatehave a unique number of antinodes for each eigenstate, starting with a single antinode for the ground state. Well, the function we have chosen (known as a *gaussian*), has only a single antinode, located at $x = 0$, so it must be the ground state!

Figure 6.5.1 – A Gaussian Wave Function in a Spring Potential Energy Well



There are a few of things to note from this result and the figure above. First, we see that in fact the ground state energy is above the bottom of the well, as we expected it to be. Second, the wave function has a built-in exponential decay in the classically forbidden region – there's no need to stitch together two different functions as we did for the finite square well with the sinusoidal and exponential functions. Third, setting the total energy of this state equal to the potential energy and solving for x gives us the classical turnaround points in terms of the particle mass and spring constant. And finally, it should be noted that unlike the previous cases, it is conventional to designate the ground state of the quantum harmonic oscillator with a zero subscript, rather than a one, for reasons that will become clear when we discuss the energy spectrum shortly.

There is still a bit of unfinished work to be done on this particular eigenstate. We found the value of the parameter α, but we do not yet have the value of A – we have not yet normalized the wave function. This is a definite integral we can just look up:

$$1 = \int_{-\infty}^{+\infty} |\psi(x)|^2 dx = A^2 \int_{-\infty}^{+\infty} e^{-2\alpha x^2} dx = A^2 \left(\sqrt{\frac{\pi}{2\alpha}}\right) \quad \Rightarrow \quad A = \left(\frac{2\alpha}{\pi}\right)^{\frac{1}{4}} \tag{6.5.8}$$

What about the other energy eigenstates? The ground state must have even symmetry about the origin, and indeed the gaussian wave function given above has this property. All the odd-numbered excited states must have odd symmetry, while all the even-numbered excited states have even symmetry (remember, the ground state is $n = 0$). It turns out that all of the excited states only differ from the ground state by multiplying the gaussian by a polynomial, known as a *Hermite polynomial*. We won't worry about how to generate these polynomials, but it is possible to understand a few of their features just using what we know about wave functions of bound particles.

First, the polynomials for odd-numbered states must include powers of x that are odd only. That way, when they multiply the symmetric gaussian, they will have the proper odd symmetry. Similarly, even-numbered states must involve only even powers of x (the ground state polynomial includes x^0, which is an even power).

Second, the number of nodes must go up by one with every increase in level. Nodes are crossings of the x-axis, and this tracks the order of the polynomial. Therefore, the Hermite polynomials must look like:

$$\begin{aligned} n = 0: &\quad H_0(x) = Ax^0 \\ n = 1: &\quad H_1(x) = Bx^1 \\ n = 2: &\quad H_2(x) = Cx^0 + Dx^2 \\ n = 3: &\quad H_3(x) = Ex^1 + Fx^3 \\ \vdots &\quad \vdots \end{aligned} \tag{6.5.9}$$

One can derive the unknown constants in these polynomials in a brute-force manner by multiplying them by the gaussian, and plugging the result into the differential equation, just as we essentially did for the ground state above. We should probably be a bit more precise about what we mean by the Hermite polynomial "multiplying the gaussian," so here is the actual normalized wave function of the n^{th} energy eigenstate in position space, in terms of $H_n(x)$:

$$\psi_n(x) = \left(\frac{\beta}{2^n n! \sqrt{\pi}}\right)^{\frac{1}{2}} H_n(\beta x) e^{-\frac{(\beta x)^2}{2}} = \frac{1}{\sqrt{2^n n!}} H_n(\beta x) \psi_o(x) \tag{6.5.10}$$

For reasons of simplicity in some other constants, we have replaced the constant α with $\frac{1}{2}\beta^2$, so:

$$\beta \equiv \sqrt{2\alpha} = \left(\frac{\kappa m}{\hbar^2}\right)^{\frac{1}{4}} \tag{6.5.11}$$

And for the sake of having a few of the lower energy eigenfunctions available to work with, here are a few of the Hermite polynomials:

$$\begin{aligned} H_o(\beta x) &= 1 \\ H_1(\beta x) &= 2\beta x \\ H_2(\beta x) &= 4\beta^2 x^2 - 2 \\ H_3(\beta x) &= 8\beta^3 x^3 - 12\beta x \end{aligned} \tag{6.5.12}$$

It's pretty obvious that $\psi_{n_1}(x)$ is orthogonal to $\psi_{n_2}(x)$ when n_1 is odd and n_2 is even, or vice-versa, since the overlap integral will be between an odd and even function and will therefore vanish. But what is truly amazing about these polynomials is that the property of *all* eigenstates being orthogonal holds, which means the integral is zero if $n_1 \neq n_2$, even when they are both odd or both even.

Energy Spectrum

If we plug $\psi_n(x)$ into Schrödinger's equation, the eigenvalues E_n come out. As complicated as the eigenfunctions and the operators in the Schrödinger equation are, the energy spectrum comes out remarkably simple:

$$E_n = \left(n+\frac{1}{2}\right)\hbar\,\omega_c, \quad n=0,1,2,\ldots, \quad \omega_c \equiv \sqrt{\frac{\kappa}{m}} = \text{angular frequency of classical oscillator} \tag{6.5.13}$$

Alert

The symbol ω_c should not be confused with the other Greek letter omega that we use in the quantum phase time-dependence: $\omega_n = \frac{E_n}{\hbar}$.

Uncertainties

We can compute uncertainties for the usual suspects (position, momentum, and energy) in the energy eigenstates in the standard way – by performing the expectation integrals and plugging them into the formula for uncertainty. But we are more clever than that. We start by noting that in the kinetic energy operator in the Schrödinger equation is quadratic in $\hat{p}$, and the potential energy operator is quadratic in $\widehat{x}$. Given the reciprocal relationship we know exists between these two quantities (think of the Fourier transform and its inverse!), it not a stretch (though we will not show it mathematically here) to claim that the average potential energy equals the average kinetic energy. This is in fact even true over a full oscillation of a mass on a spring in classical physics. Given this, we can take some shortcuts.

Using the fact that the expectation value of the total energy for a given energy eigenstate is simply the energy eigenvalue, we can deduce that the average kinetic and potential energies are half the energy eigenvalue:

$$E_n = \langle E\rangle_n = \langle KE+PE\rangle_n = \langle KE\rangle_n + \langle PE\rangle_n \quad \Rightarrow \quad \langle KE\rangle_n = \langle PE\rangle_n = \frac{1}{2}E_n = \frac{1}{2}\left(n+\frac{1}{2}\right)\hbar\,\omega_c \tag{6.5.14}$$

We can carry this result into finding the uncertainty in position and momentum as well. Start by noting that symmetry demands that the expectation value of position and momentum are both zero, since the probability density (in both position and momentum space) is symmetric about the origin. This means that the uncertainty in this values depends only upon the expectation value of their squares. But these squares are proportional to the potential and kinetic energies, so we get answers without ever performing a gaussian integral:

$$\begin{aligned} \langle x^2\rangle_n &= \frac{2}{\kappa}\langle PE\rangle_n = \left(n+\tfrac{1}{2}\right)\hbar\left(\frac{\omega_c}{\kappa}\right) = \left(n+\tfrac{1}{2}\right)\hbar\frac{1}{\sqrt{\kappa m}} \\ \langle p^2\rangle_n &= 2m\langle KE\rangle_n = \left(n+\tfrac{1}{2}\right)\hbar\,(m\omega_c) = \left(n+\tfrac{1}{2}\right)\hbar\sqrt{\kappa m} \end{aligned} \tag{6.5.15}$$

Plugging into the uncertainty equations:

$$\begin{aligned} \Delta x &= \sqrt{\langle x^2\rangle} = \sqrt{\left(n+\tfrac{1}{2}\right)\hbar}\;(\kappa m)^{-\frac{1}{4}} \\ \Delta p &= \sqrt{\langle p^2\rangle} = \sqrt{\left(n+\tfrac{1}{2}\right)\hbar}\;(\kappa m)^{\frac{1}{4}} \end{aligned} \tag{6.5.16}$$

Whenever we have the uncertainties for position and momentum, it is natural to want to test the uncertainty principle. So multiplying these together gives:

$$\Delta x\Delta p = \left(n+\frac{1}{2}\right)\hbar \tag{6.5.17}$$

The uncertainties both get bigger as the energy level goes up, so the ground state represents the smallest value of this product, and it turns out that the ground state of the harmonic oscillator ($n=0$) provides the very limit of the uncertainty principle!

Why This Potential?

It is natural to ask why we are studying this potential at all. After all, quantum particles are not attached to each other by tiny springs. Is this just an exercise to solve a problem with no practical application? Not at all! In fact this is probably the *most* applicable of the models we look at in introductory quantum theory. The reason is that when particles are bound to each other, the potential energy curve forms a well that is quite similar to that of a spring potential. We actually covered this fact already in Physics 9HA, when we discussed modeling particle bonds as springs. We can use this process to estimate the energy spectrum for bonds between particles for which we have a good idea of the potential energy function. We simply find the equivalent spring constant for the bond in question, call that value "κ", and use the results that we derived here.

6.6: The Bohr Model of the Hydrogen Atom

The Classical Hydrogen Atom

Every quantum mechanical model we have discussed so far has been built from a classical system that we are familiar with. A square well is a pair of walls between which a ball is bouncing, and a harmonic oscillator is a mass on a spring. Here we will once again start with a classical model originated by Ernest Rutherford and later modified for quantum physics by Niels Bohr. It should be noted before continuing that Bohr did this work over a decade before Schrödinger introduced the equation that bears his name, so appropriate treatment of "matter waves" was not yet known.

The classical model of Rutherford perfectly parallels what we see in our solar system. Heavy, electrically positively-charged protons occupy a small nucleus, and they attract the much lighter negatively-charged electrons, and the combination of these form an atom. Just as gravity doesn't cause the solar system to collapse, the electrical attraction between the protons do not cause atoms to collapse – the tangential orbital motion prevents it.

The orbital model of the atom is even close to that of the solar system mathematically. It had been known for over a century that the electrical force between point-like charges obeys an inverse-square law, just like that of Newton's law of universal gravitation (see Equation 7.1.3 in the 9HA textbook):

$$\vec{F}_{electric} = \frac{kq_1q_2}{r^2}\,\hat{r} \tag{6.6.1}$$

The q's are electric charge, which can be either positive or negative. When the two charges have opposite signs (as in the case of protons and electrons), the direction is $-\hat{r}$, which means the force is attractive. This is known as *Coulomb's Law*, and the SI units of electric charge (C) bear Coulomb's name. While the constant k that plays the role of G from gravitation is pretty common to see, this is the one and only time we will use it, as we don't want to confuse it with wave number. Instead, we will exchange it for another constant that is even more ubiquitous in the subject of electricity. This advantage of using this particular constant will have to remain a mystery for now (you'll see it in Physics 9HD!):

$$k \equiv \frac{1}{4\pi\epsilon_o} \quad \Rightarrow \quad \vec{F}_{electric} = \frac{q_1q_2}{4\pi\epsilon_o r^2}\,\hat{r}\,, \quad \epsilon_o = 8.85\times10^{-12}\,\frac{C^2}{Nm^2} \tag{6.6.2}$$

Of course, if we are going to somehow link this to quantum mechanics, we will be better off with an energy approach, and as in the case of gravitation (Equation 7.3.6, Physics 9HA textbook), this force can be expressed as a potential energy that is only a function of the particle separation:

$$V(r) = \frac{q_1q_2}{4\pi\epsilon_o r} \tag{6.6.3}$$

Note that this potential does not have an overall minus sign as the gravitational potential does, because the charges carry signs with them. When they are opposite (giving an overall negative value), the force is attractive, as it is in the case of gravity where both masses are always positive.

This force is a central force, so it is conservative, and we can apply conservation of mechanical energy to this system. Before we do, there is one more simplification to make. It turns out that while the mass of the proton and electron are far apart, they have precisely the same magnitude of charge, usually referred to as "e". Rutherford's nucleus allowed for many protons, and this integer is typically represented with a "Z". We are interested in the effect this nucleus has on a single electron (at least as a start), so the atomic potential for a single electron is written as:

$$V(r) = -\frac{Ze^2}{4\pi\epsilon_o r} \tag{6.6.4}$$

Now at last we can express the total energy of the electron as it orbits the nucleus:

$$E_{tot} = \frac{1}{2}mv^2 - \frac{Ze^2}{4\pi\epsilon_o r} \tag{6.6.5}$$

As with the case of gravitation, in the absence of outside forces, the closer the electron gets to the nucleus the faster it moves, as its potential energy goes down (it becomes more negative at smaller values of r).

Lastly, it should be noted that we don't have to worry too much about treating the nucleus as a fixed point. Usually we solve such two-body problems using something called "reduced mass" (which we first encounters in Physics 9HA in Equation 8.2.12), but in this case the proton is nearly 2000 times as massive as the electron, so our approximation of the proton being a fixed central source of the electrical force is a good one.

Two Puzzles

As nice and tidy as this classical picture of the atom is, it runs into two major flaws when the quantum theory is taken into account. It was known since the time of Maxwell that light (and all EM radiation) is produced when electric charges accelerate. Rutherford also showed that atoms had a localized, hard nucleus, which meant that atoms consisted of negatively-charged electrons bound by the electric force to a positively-charged nucleus, like satellites are bound in their orbits by gravitational forces. But this posed a conundrum – electric charges in orbits are accelerating centripetally, so they should constantly be radiating, and since light carries away energy, the electrons should spiral down into the nucleus, which it clearly does not.

The second puzzle is characteristically quantum-mechanical: There is no reason to believe that any light frequencies released by atoms that lose energy should be preferred over any others. But it was known for some time that viewing light emitted by various elements through a diffraction grating reveals distinct sets of *spectral lines* – lines that are separated according to frequency.

With what little quantum theory we have already learned, we have insight into both of these phenomena, since an electron is clearly in a bound state due to the electrical potential. The first puzzle is "solved" by remembering that the ground state of a particle in a bound state can never be at the bottom of the well. It's spiral down forever until it has no energy because there is no zero-energy eigenstate available. The second puzzle is just another example of what we have seen about bound states generally being quantized.

But as noted above, Bohr did not have Schrödinger's equation, and he set out to stitch together the ideas of Rutherford's classical picture of the orbiting electron atom with the new paradigm of matter waves exhibiting de Broglie wavelengths.

Bohr's Model

Bohr reasoned that a particle whose location is extended in space in the form of a wave and is orbiting a central point would have to interfere with itself when it gets all the way around. For the particle to remain in such an orbit, it would have to (as we call it now) "match boundary conditions" – its wave function would have to be in phase with itself when a full orbit is completed. He then had to address the issue of orbits of inverse-square forces being elliptical. While he was not able to motivate *why*, he postulated that the orbits can only be circular. While this assumption makes little sense, its results are striking enough to let it slide, and maybe take it up later.

The in-phase and circular orbit requirements, along with de Broglie's formula for the wavelength of a matter wave, and the coulomb potential for two point charges is all that was required to complete this model. Let's see what this semi-classical approach predicts for the simplest of atoms – hydrogen...

The assumption that the electron follows a circular orbit requires that this force causes a centripetal acceleration, so calling the mass of the electron "m_e", we have:

$$m_e\frac{v^2}{r}=\frac{e^2}{4\pi\epsilon_o r^2}\quad\Rightarrow\quad p^2=(m_e v)^2=\frac{m_e e^2}{4\pi\epsilon_o r} \tag{6.6.6}$$

We put this in terms of the magnitude of the electron's momentum, so that we can use the deBroglie relation. Applying Bohr's assumption that an integer number of full wavelengths fit within the circular orbit (so that it meets itself in phase) gives:

$$\left.\begin{array}{ll} deBroglie: & p=\dfrac{h}{\lambda} \\ Bohr: & 2\pi r=n\lambda \end{array}\right\}\quad\Rightarrow\quad p^2=\left(\frac{nh}{2\pi r}\right)^2=\frac{n^2\hbar^2}{r^2} \tag{6.6.7}$$

Plugging this back into Equation 6.6.6 and solving for the radius of the orbit gives:

$$r_n=n^2\left(\frac{4\pi\epsilon_o\hbar^2}{m_e e^2}\right)\equiv n^2 a_o \tag{6.6.8}$$

The orbit radii are thus quantized! The quantity a_o, derived purely from physical constants, is called the *Bohr radius*, which is a useful unit of length measurement, even in the more enlightened quantum mechanical model that came later.

Of course, quantization of the orbital radii was not what we were after – we are interested in the energy levels of the hydrogen atom, which are somehow involved in providing energy to emitted photons, and we want to know why the hydrogen atom doesn't radiate away all its energy. To this end, we note that the assumption of a circular orbit results in a very simple relationship between the kinetic, potential, and total energy. Dividing Equation 6.6.6 by twice the mass of the electron gives us its kinetic energy, and comparing the result with Equation 6.6.4 (with $Z=1$ for the hydrogen atom) gives:

$$\frac{p^2}{2m}=\frac{r}{2}\left(\frac{e^2}{4\pi\epsilon_o r^2}\right) \quad \Rightarrow \quad KE=-\frac{1}{2}PE \quad \Rightarrow \quad E_{tot}=KE+PE=\frac{1}{2}PE=-\frac{e^2}{8\pi\epsilon_o r} \tag{6.6.9}$$

But the radii are quantized, so plugging this in gives quantized energy levels:

$$E_n=-\frac{e^2}{8\pi\epsilon_o r_n}=-\frac{1}{n^2}\left(\frac{m_e e^4}{2(4\pi\epsilon_o)^2\hbar^2}\right) \tag{6.6.10}$$

The energy spectrum is also quantized. The quantity in parentheses is a unit of energy known as a *Rydberg*, which $\approx 13.6eV$.

Bohr reasoned that the hydrogen atom doesn't radiate away all of its energy, because the lowest energy level (which we would now call the ground state) corresponds to exactly one wavelength fitting in the orbit, so $n=1$ is the lowest it can go.

Emission/Absorption Spectrum

The problem of only seeing certain spectral lines is also solved in this model, if one insists that the atom can only exist in one of these quantized states, and therefore the only energy transitions it can make are between the allowed energy levels. A transition that lowers the energy level of the hydrogen atom from \(n_1) to \(n_2) frees up the amount of energy equal to the difference, which then goes into an emitted photon according to Planck's relation:

$$hf=\Delta E=E_{n_1}-E_{n_2}=-13.6eV\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right) \tag{6.6.11}$$

This matched perfectly with experiment! So although there are many problems with this model, a more evolved version will need to agree with this energy spectrum. Note that this quantized energy change works both ways – when a hydrogen atom absorbs a photon, it must absorb an amount of energy that carries the atom from one of its quantized energy levels to another (higher) one.

So Many Flaws!

The final result is so striking that to this very day we memorialize the result in textbooks like this one, but this euphoria only lasts so long, and eventually we have to address the obvious flaws of the model. The most obvious is probably the arbitrary assumption of circular orbits. If we allow for elliptical orbits as we see in gravitation, we once again get a continuum of energies possible, even with the requirement that the orbiting matter wave is in phase with itself. But there are other flaws as well, some of which come from disagreement with experiment, and others just don't agree with what we already know about quantum theory.

Orbital motions of planets and electrons share the property that their orbits remain in a plane. This, along with the circular orbits assumption, is why this model fits into the "One-Dimensional Models" chapter. But this causes a very large problem with the uncertainty principle. If we call the plane of orbit of the electron the $x-y$ plane, then this model confines the electron to a single, precise value of z. Its position remains undetermined in its orbital plane, but we exactly know its z-position. This means that we know *nothing* about its z-component of momentum, which really messes-up what we thought we know about its kinetic energy.

Another problem comes from experiments. There are many of these, but the most obvious comes from what happens when hydrogen atoms are sent through magnetic fields. A very basic result from electromagnetism is that charges that go in circles behave like magnets – they can be deflected by non-uniform magnetic fields as they pass through them (we will discuss this in a future chapter). Well, there is no way for a Bohr atom to *not* have this magnetic property, but we do see hydrogen atoms (most notably, all of them in their ground states) pass through magnetic fields without deflecting. Well, okay, they still do, but that cannot be explained from the orbital motion of the electron – again, this is coming attractions.

CHAPTER OVERVIEW

7: Quantum Theory in Three Dimensions

7.1: Schrödinger's Equation in 3-Dimensions

Adding Degrees of Freedom

When we extend our work on quantum mechanics in one dimension to three dimensions, we have to take into account what happens to how we describe quantum states. The main difference that arises is that we still need to get to a single number (probability) from three times as many dimensions as before. Instead of describing the wave function in terms of a single number x, we require three numbers, x, y, and z. For cases when we want to remain ambiguous about the coordinate system in use, we will replace explicit use of variables like (x, y, z) with the more generic position vector $\vec{r}$.

$$\Psi(x, y, z, t) \rightarrow \Psi(\vec{r}, t) \tag{7.1.1}$$

As before, the magnitude-squared of the wave function is a probability density, but in three dimensions, this becomes a volume density instead of a line density. Computations of probabilities in three dimensions requires integration over a three dimensional space:

$$\text{probability that particle is found in a tiny volume } dV \text{ located at position } \vec{r} = |\Psi(\vec{r}, t)|^2 dV \tag{7.1.2}$$

The tiny volume element dV takes on different forms depending upon the coordinate system used. In this class, we will only be using Cartesian and spherical coordinates:

$$\begin{array}{rcll} & & \text{Cartesian} & \text{spherical} \\ dV & = & dx\ dy\ dz & r^2\ dr\ \sin\theta\ d\theta\ d\phi \end{array} \tag{7.1.3}$$

Naturally the normalization condition is:

$$\int\limits_{all\ space} |\Psi(\vec{r}, t)|^2 dV = 1 \tag{7.1.4}$$

Written out explicitly for Cartesian and spherical coordinates, this is:

$$\begin{aligned} &\int\limits_{-\infty}^{\infty} dx \int\limits_{-\infty}^{\infty} dy \int\limits_{-\infty}^{\infty} dz\ |\Psi(x, y, z, t)|^2 = 1 \\ &\int\limits_{0}^{\infty} r^2\ dr \int\limits_{0}^{\pi} \sin\theta\ d\theta \int\limits_{0}^{2\pi} d\phi\ |\Psi(r, \theta, \phi, t)|^2 = 1 \end{aligned} \tag{7.1.5}$$

As before, we can use eigenstates of momentum (plane waves) as our "unit vectors," and the relationship between the y and z components of momentum with the y and z coordinates is exactly the same as it was for the case of x in one dimension, which means that position space and momentum space wave functions are related through a Fourier transform, which now must be performed in all three directions.

One interesting effect arises from the extension to three dimensions. In the case of one dimension, the stationary-state wave function was completely defined by a single number x, the position along the one dimension. The wave function also contained all of the information about the stationary state, namely the energy. This is also reversible – if we know the energy of the particle (say we know that a particle-in-a-box is in its ground state), then we also have its full wave function and all of the information about its state.

When we get to three dimensions, however, knowing the wave function requires knowing the three numbers that define position. These three numbers will provide all the information about the particle's state, including its energy, but this time it is not reversible – knowing the energy (a single number) cannot possibly tell us everything about the state of the particle. There has to be two other such numbers. If the particle is bound, these numbers will – like the energy – be quantized, and we will label them with integers as we already do with the energy. These integer labels are called *quantum numbers*, and are frequently used to label eigenstates in the same way that n labeled the energy eigenstates in one dimension.

Schrödinger's Equation

Schrödinger's equation is easy to expand to three dimensions. All that is required is to use the kinetic and potential energy operators in three dimensions in the Hamiltonian. In Cartesian coordinates, we have:

$$\widehat{H}=\frac{1}{2m}\left(\hat{p}_x^2+\hat{p}_y^2+\hat{p}_z^2\right)+\widehat{V}=-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V\left(x,y,z\right) \tag{7.1.6}$$

We can write this without reference to a choice of coordinate system by replacing the set of second derivatives with the Laplacian operator (which we can look up for whatever coordinate system we like):

$$\widehat{H}=-\frac{\hbar^2}{2m}\nabla^2+V\left(\vec{r}\right) \tag{7.1.7}$$

Once again, we are faced with a partial differential equation, this time in *four* variables. Just as before, we can use the separation of variables method to select-out the stationary-state subset of solutions:

$$\Psi\left(\vec{r},t\right)=\psi\left(\vec{r}\right)e^{-i\omega t},\qquad where:\quad -\frac{\hbar^2}{2m}\nabla^2\psi\left(\vec{r}\right)+V\left(\vec{r}\right)\psi\left(\vec{r}\right)=E\,\psi\left(\vec{r}\right),\quad \omega=\frac{E}{\hbar} \tag{7.1.8}$$

Okay, now we have a differential equation with three variables to solve. If the separation of variables trick is so effective, why not just use it again? Not so fast! Unlike separating the time variable from the spatial variables, separating the spatial variables from each other is trickier business, because we can choose any set of coordinates we like. That is, there is nothing about the Schrödinger equation written above in Cartesian coordinates that is any more "correct" than if it is written in spherical coordinates, so how do we know whether we can (or should) do a separation of variables in Cartesian coordinates?

The answer is utility – we trust that like the case of separation of the time variable, we will be able to construct more general solutions from linear combinations of whatever solutions we arrive at after separating variables, so we are free to try the separation in any coordinate system. The system we choose should be the one that is easiest to work with for the physical situation given. If the potential is easy to work with in a specific coordinate system (typically due to some spatial symmetry inherent in the potential), then that is the one to use. Ultimately whatever separation we choose will lead to its own set of three quantum numbers, as we will see with our two cases of Cartesian and spherical coordinates.

Separation of Variables in Cartesian Coordinates

Potentials with particular properties encourage us to separate variables in the Cartesian coordinate system, so let's look at how this works. We seek solutions to the stationary-state Schrödinger equation that admit wave functions which can be written as a product of three functions of single variables:

$$\psi\left(x,y,z\right)=X\left(x\right)Y\left(y\right)Z\left(z\right) \tag{7.1.9}$$

Plugging this into the stationary-state Schrödinger equation, and dividing the whole equation by the wave function separates it into terms that are functions of only x, y, and z, along with a potential that so far we have not restricted:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)X\left(x\right)Y\left(y\right)Z\left(z\right)+V\left(x,y,z\right)X\left(x\right)Y\left(y\right)Z\left(z\right)=E\,X\left(x\right)Y\left(y\right)Z\left(z\right) \tag{7.1.10}$$

$$\Rightarrow\quad \frac{1}{X}\frac{\partial^2 X}{\partial x^2}+\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2}+\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2}-\frac{2m}{\hbar^2}V\left(x,y,z\right)=-\frac{2mE}{\hbar^2}$$

For this method to be useful, we need to be able to separate the entire equation into a sum of terms that are exclusively functions of one variable at a time (x, y, or z). To see how this works, consider the following:

$$f\left(x\right)+g\left(y\right)+h\left(z\right)=constant \tag{7.1.11}$$

We can plug any x value we like into $f\left(x\right)$, without changing any of the y or z values. For this equation to remain correct, it must mean that $f\left(x\right)$ is the same constant value for all choices of x. The same argument can be made for $g\left(y\right)$ and $h\left(z\right)$. This gives us three separate equations, each in a single variable.

The first three terms are already separated into x, y, and z, but the potential function poses a problem. This method is only really effective in two cases: When the potential is a constant (universally, or piecewise), or when it splits up into a sum of functions of

single variables. We will look at examples of each of these cases, starting with the simplest – where there is essentially no potential function at all.

Free Particle

If the potential is universally constant, then the particle is obviously free. As with the one-dimensional free particle, the stationary-state Schrödinger's equation gives us wave functions of energy eigenstates, but we can filter-out the momentum eigenstates (plane waves) if we wish. Plugging zero in for the potential allows us to separate the equation into three differential equations in single variables. The choices for the form of the constants below will become quickly apparent:

$$\frac{1}{X}\frac{d^2X}{dx^2} = -k_x^2, \qquad \frac{1}{Y}\frac{d^2Y}{dy^2} = -k_y^2, \qquad \frac{1}{Z}\frac{d^2Z}{dz^2} = -k_z^2, \qquad k_x^2 + k_y^2 + k_z^2 = \frac{2mE}{\hbar^2} \tag{7.1.12}$$

The plane wave solutions of each of the separate differential equations are:

$$X(x) = A_x e^{ik_x x}, \qquad Y(y) = A_y e^{ik_y y}, \qquad Z(z) = A_z e^{ik_z z} \tag{7.1.13}$$

Reconstructing the full momentum eigenstate wave function, we get:

$$\psi(x,y,z) = X(x)\,Y(y)\,Z(z) = Ae^{i(k_x x + k_y y + k_z z)} \tag{7.1.14}$$

If we define the momentum vector $\vec{p}$ in terms of a wave vector $\vec{k} = k_x\hat{i} + k_y\hat{j} + k_z\hat{k}$, we get the rather compact plane wave solution moving in a specific direction:

$$\psi(\vec{r}) = Ae^{i\vec{k}\cdot\vec{r}}, \qquad \vec{k} = \frac{\vec{p}}{\hbar}, \qquad E = \frac{p^2}{2m} = \frac{\hbar^2}{2m}\left(\vec{k}\cdot\vec{k}\right) \tag{7.1.15}$$

This plane wave is also an energy eigenstate, so the full time-dependent wave function can also be written:

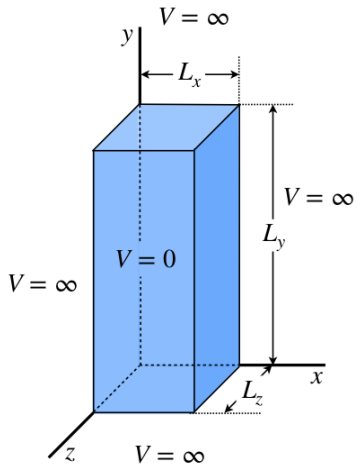
$$\psi(\vec{r},t) = Ae^{i\left(\vec{k}\cdot\vec{r} - \omega t\right)}, \qquad \omega = \frac{E}{\hbar} \tag{7.1.16}$$

7.2: 3-Dimensional Models

Particle in a 3D Infinite Square Well

The infinite square well in three dimensions has the same property as the one-dimensional box – the potential is zero everywhere inside, and instantly becomes infinity at the boundaries. The one-dimensional case had a specified length, but we will not saddle this infinite well with the same width in all three directions, meaning we will confine the particle to a rectangular prism, not a cube. We will define our coordinate system so that the walls are parallel to the three planes, and we will place the origin at one of the box corners. The lengths of the walls along the x, y, and z axes we will call L_x, L_y, L_z, respectively.

Figure 7.2.1 Three-Dimensional Infinite Square Well



Mathematically, the potential is written:

$$V(x,y,z) = \begin{cases} 0 & 0 < x < L_x \ \ and \ \ 0 < y < L_y \ \ and \ \ 0 < z < L_z \\ \infty & \text{elsewhere} \end{cases} \tag{7.2.1}$$

As we saw for the one-dimensional box, we can use a combination of two oppositely-moving plane waves (for each of the three axes) to construct a wave function of definite energy that vanishes at the walls (i.e. sinusoidal functions). The individual solutions to the differential equations in x, y, and z are the same as before, with two exceptions: Each dimension involves a separate harmonic number n, so like the 1-dimensional well, all the wave functions are sines:

$$X(x) = A_x \sin\frac{n_x \pi\, x}{L_x} \qquad Y(y) = A_y \sin\frac{n_y \pi\, y}{L_y} \qquad Z(z) = A_z \sin\frac{n_z \pi\, z}{L_z}, \qquad n_x,\ n_y,\ n_z = 1,\ 2,\ \ldots \tag{7.2.2}$$

Before we move on to the energy spectrum, let's construct the spatial full wave function by multiplying the partial wave functions. We also have to deal with normalizing the full wave function. Normalization does not tell us anything about the values of A_x, A_y, and A_z, but their product must equal the normalization constant for the full wave function. The normalization integral is over all three dimensions and integrals over x will only affect $X(x)$, and similarly for y and z, so the normalization constant for the full wave function turns out to be the same as the product of the normalization constants for the three separate one-dimensional wave functions.

$$\psi_{n_x n_y n_z}(x,y,z) = \sqrt{\frac{8}{L_x L_y L_z}} \sin\frac{n_x \pi\, x}{L_x} \sin\frac{n_y \pi\, y}{L_y} \sin\frac{n_z \pi\, z}{L_z}, \qquad n_x,\ n_y,\ n_z = 1,\ 2,\ \ldots \tag{7.2.3}$$

In keeping with our notation of labeling the wave function with the quantum numbers, we have labeled the energy eigenstate wave function accordingly.

Plugging the wave function back into the stationary-state Schrödinger equation, we get the following energy spectrum:

$$E_{n_x n_y n_z} = \left(\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right)\frac{h^2}{8m}, \quad n_x,\ n_y,\ n_z = 1,\ 2,\ \ldots \tag{7.2.4}$$

Alert

One might be tempted to think that the ground state energy of this particle occurs when the n along the longest dimension is 1, while the others are zero, but It should be emphasized that the minimum value of all three n values is 1. None of the three modes can provide a zero contribution to the energy.

Another way that the three-dimensional case differs from the one-dimensional case is apparent if we consider the hierarchy of the energy spectrum. Suppose for example, we wanted to draw an energy-level diagram for this spectrum. We know that ψ_{111} is the ground state, but which quantum state would be the first excited state? The answer depends upon the dimensions of the well. If the L_x is greater than the other two box dimensions, then the smallest increase in the total energy will come from incrementing n_x from 1 to 2, and the first excited state would be ψ_{211}. What about the second excited state? Well, now we need even more information. If L_x is only slightly greater than L_y dimension (and both are longer than L_z), then the second excited state would be ψ_{221}. But if L_x is significantly longer than the other dimensions, then the second excited state would be ψ_{311}. In other words, with three quantum numbers, we have lost the ability (at least for this case) to express the energy levels with a single integer.

The 3D Harmonic Oscillator

As our final example of a potential that allows for separation of variables in Cartesian coordinates, we consider the three dimensional harmonic oscillator, which has a potential that is a sum of functions purely of x, y, and z. In general, the spring constants are different for each direction, so:

$$V(x,y,z) = \frac{1}{2}\kappa_x x^2 + \frac{1}{2}\kappa_y y^2 + \frac{1}{2}\kappa_z z^2 \tag{7.2.5}$$

This sort of model might be useful for crystal lattices where the bonds differ along the different dimensions. Plugging this into Equation 7.1.10 results in an equation with three separated functions again:

$$\left[\frac{1}{X}\frac{\partial^2 X}{\partial x^2} + \frac{\kappa_x m}{\hbar^2}x^2\right] + \left[\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} + \frac{\kappa_y m}{\hbar^2}y^2\right] + \left[\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} + \frac{\kappa_z m}{\hbar^2}z^2\right] = -\frac{2mE}{\hbar^2} \tag{7.2.6}$$

Following the same procedure as before gives us three separate differential equations. This decoupling maneuver once again leaves us with three wave function pieces, which are multiplied together to get the full wave function. As with the case of the square well, the energy contributions of the partial wave functions are added together to give the total energy of the state:

$$E_{n_x n_y n_z} = \left(n_x + \frac{1}{2}\right)\hbar\sqrt{\frac{\kappa_x}{m}} + \left(n_y + \frac{1}{2}\right)\hbar\sqrt{\frac{\kappa_y}{m}} + \left(n_z + \frac{1}{2}\right)\hbar\sqrt{\frac{\kappa_z}{m}} \tag{7.2.7}$$

This gives us a nice way to describe bonds along different axes in general, but of particular interest is the *isotropic harmonic oscillator*, which involves equal spring constants (κ) in all three directions. In this case, the energy spectrum reduces to:

$$E_{n_x n_y n_z} = \left(n_x + n_y + n_z + \frac{3}{2}\right)\hbar\omega_c\,, \quad \omega_c = \sqrt{\frac{\kappa}{m}} \tag{7.2.8}$$

Notice that even with the simplification of isotropy, three quantum numbers are required to define the state.

Degeneracy

Notice that unlike one-dimensional potentials, in these cases a single quantum number does not define the energy. But there is even more to it than that. Looking at the case of the three-dimensional box again, suppose it has three equal sides: $L_x = L_y = L_z = L$. In this case, there exist three distinct quantum states that possess the same total energy, namely ψ_{211}, ψ_{121}, and ψ_{112}. These states clearly possess equal energies, and yet they are distinct because, for example, the states ψ_{211} and ψ_{121} yield different uncertainties in the x-component of the particle's position (ψ_{211} has two antinodes along the x axis, while ψ_{121} has only one). A similar thing occurs (only more dramatically) for the isotropic harmonic oscillator, as any combination of n_x, n_y, and n_z that gives the same sum will result in the same energy.

When multiple quantum states yield the same energy, they are said to be *degenerate*, and if there are a total of j distinct states for the same energy, that energy level is said to be *j-fold degenerate*. Typically degeneracy comes about due to obvious symmetries, such as in the cases mentioned above. All we need to do is rename our axes, and the states morph into each other, so naturally the energies are the same. But occasionally degeneracies arise unexpectedly, through what can only really be described as a coincidence. These are called *accidental degeneracies*. An example of one of these for the three-dimensional square well with all

three sides of equal length arises for the states ψ_{511}, ψ_{151}, ψ_{115}, and ψ_{333} – this energy level is 4-fold degenerate, rather than the "expected" 3-fold degeneracy. Naturally the first three states are not unexpectedly degenerate, but the fourth seems to come from out of nowhere.

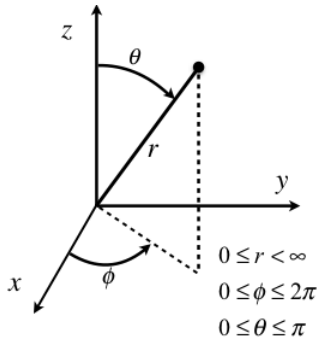
Symmetric quantum systems are common in physics, and degeneracy follows them everywhere. This can cause difficulty in developing theory, as some internal structure can be obscured when different configurations result in the same energy spectrum. The trick then is to introduce an external perturbation that breaks the symmetry, thereby separating otherwise degenerate states. The analogous case for the cubical box would be squeezing or stretching one of the dimensions slightly. This also can work in the other direction – we might see unexpected additional spectral lines that indicate that there is additional structure present that breaks the symmetry we thought existed. So the analogous case for this is an infinite well that we think should be cubical, but provides a spectrum with energy levels landing between those that we compute.

7.3: Central Forces

Separation of Variables

Working in spherical coordinates is significantly more difficult than working in Cartesian coordinates. So why do it? Because point-like particles are sources for spherically-symmetric potentials that affect other particles. Understanding how to work in spherical coordinates is essential for solving the hydrogen atom in particular. First a reminder of the coordinates themselves:

Figure 7.3.1 – Spherical Coordinates



We begin with the stationary-state Schrödinger equation in three dimensions. We need to write the Laplacian operator found in the Hamiltonian (Equation 7.1.7) in spherical coordinates. This we can simply look up (deriving it from the transformation between coordinate systems is no fun):

$$\nabla^2 = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \left(\frac{1}{\sin\theta}\right)\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \left(\frac{1}{\sin^2\theta}\right)\frac{\partial^2}{\partial\phi^2}\right] \tag{7.3.1}$$

The stationary-state Schrodinger equation in spherical coordinates is therefore:

$$-\frac{\hbar^2}{2mr^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \left(\frac{1}{\sin\theta}\right)\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \left(\frac{1}{\sin^2\theta}\right)\frac{\partial^2}{\partial\phi^2}\right]\psi(r,\theta,\phi) + V(r,\theta,\phi)\psi(r,\theta,\phi) = E\psi(r,\theta,\phi) \tag{7.3.2}$$

Our goal is to collect similar variables, and before we use our usual trick, we can get a head start on that by removing the r^{-2} factor from the second and third terms. Collecting all the terms with factors of r on one side of the equation then gives:

$$\left(\frac{1}{\sin\theta}\right)\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)\psi(r,\theta,\phi) + \left(\frac{1}{\sin^2\theta}\right)\frac{\partial^2}{\partial\phi^2}\psi(r,\theta,\phi) = \frac{2mr^2}{\hbar^2}[V(r,\theta,\phi) - E]\psi(r,\theta,\phi) - \frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)\psi(r,\theta,\phi) \tag{7.3.3}$$

Now follow our two-step separation trick: Write the wave function as a product of single-variable functions...

$$\psi(r,\theta,\phi) = R(r)\Theta(\theta)\Phi(\phi) \tag{7.3.4}$$

...and divide the whole equation by the wave function:

$$\frac{1}{\Theta}\left(\frac{1}{\sin\theta}\right)\frac{d}{d\theta}\left(\sin\theta\frac{d}{d\theta}\right)\Theta + \frac{1}{\Phi}\left(\frac{1}{\sin^2\theta}\right)\frac{d^2}{d\phi^2}\Phi = \frac{2mr^2}{\hbar^2}[V(r,\theta,\phi) - E] - \frac{1}{R}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)R \tag{7.3.5}$$

We haven't quite separated the angular variables yet, but let's come back to that in a bit. First, we note that the only thing keeping the right-hand side of the equation from containing only functions of r is the potential. This method therefore only helps us when the potential is a function of r, which is the case for point-source potentials, as discussed above. With the left-hand side only a function of the angular coordinates, and the right hand side only functions of r, they must both equal a common constant (which we will for the time being call "C_1"), giving us (after a bit of rearranging) what is known as the *radial equation*:

$$-\frac{\hbar^2}{2mr^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)R(r) + \frac{\hbar^2 C_1}{2mr^2}R(r) + V(r)R(r) = E\,R(r) \tag{7.3.6}$$

Plugging C_1 into the right-hand side of Equation 7.3.5, we now set out to separate the angular functions:

$$\frac{1}{\Theta}\left(\frac{1}{\sin\theta}\right)\frac{d}{d\theta}\left(\sin\theta\frac{d}{d\theta}\right)\Theta+\frac{1}{\Phi}\left(\frac{1}{\sin^2\theta}\right)\frac{d^2}{d\phi^2}\Phi=C_1 \tag{7.3.7}$$

Multiply the equation by $\sin^2\theta$ and collect the functions of each variable to get:

$$\frac{1}{\Theta}\left[\sin\theta\frac{d}{d\theta}\left(\sin\theta\frac{d}{d\theta}\right)\Theta-C_1\sin^2\theta\right]=-\frac{1}{\Phi}\frac{d^2}{d\phi^2}\Phi \tag{7.3.8}$$

Separated at last, we can state that both sides of this equation equal a constant (which will will call "C_2"), giving us two more differential equations in single variables, the *polar equation* (in θ) and the *asimuthal equation* (in ϕ):

$$\sin\theta\frac{d}{d\theta}\left(\sin\theta\frac{d}{d\theta}\right)\Theta(\theta)-\left(C_1\sin^2\theta+C_2\right)\Theta(\theta)=0 \tag{7.3.9}$$

$$\frac{d^2}{d\phi^2}\Phi(\phi)+C_2\Phi(\phi)=0 \tag{7.3.10}$$

While the solution to the radial equation depends upon the details of the potential, the solutions to the polar and azimuthal equations do not (other than the fact that the potential is strictly a function of r). We have seen the azimuthal differential equation before, and we can write the solution this way:

$$\Phi(\phi)=Ae^{\pm i\sqrt{C_2}\phi} \tag{7.3.11}$$

We cannot go any further to determine constants in this solution without boundary conditions. In our previous encounters with this differential equation, we found our boundary conditions from the potential (e.g. the wave function must vanish at the walls of the infinite square well). That is not the case here, but we *do* have a condition we can impose: If r and θ are held fixed and we change ϕ by some multiple of 2π, then we come back to the same point in space. We therefore insist that the function Φ returns to the same value every time ϕ changes by some multiple of 2π. This is ensured by insisting that the constant $\sqrt{C_2}$ is an integer. For reasons that will become clear later (and with the ancillary benefit of not confusing it with the mass of the particle), we name this integer "m_l," giving us:

$$\Phi_{m_l}(\phi)=Ae^{im_l\phi},\qquad m_l=0,\ \pm1,\ \pm2,\ \ldots \tag{7.3.12}$$

The solution to the polar equation, while it doesn't depend upon the potential, is no picnic, and we will not delve into that bit of mathematics. Its solution is called a *Legendre polynomial*, and like the azimuthal case, there is a periodic boundary condition involved, which once again introduces an integer-valued variable, this time related to the constant C_1, which we call "l." Specifically, it turns out:

$$C_1=-l(l+1),\qquad l=0,\ 1,\ 2,\ \ldots \tag{7.3.13}$$

Notice that the polar equation also includes the constant C_2, which means its solution also depends upon m_l. Indeed, this links the integers l and m_l intimately, which explains the subscript on m_l. For this reason, the functions $\Theta_{lm_l}(\theta)$ and $\Phi_{m_l}(\phi)$ are usually thrust together to make a single function (called *spherical harmonic* functions, which have been solved, and can simply be looked-up) of both variables, and this new function includes both quantum numbers (typically the "l" subscript in m_l is suppressed in the variable representing this function, as the relationship between the two quantum numbers m and l is understood, and in this context m is not going to be confused with the mass):

$$Y_{lm}(\theta,\phi)=\Theta_{lm_l}(\theta)\Phi_{m_l}(\phi) \tag{7.3.14}$$

[*Note: In the literature, the indices defining spherical harmonics are frequently written with the m_l index raised: Y_l^m.*]

For a given value of l, it so happens that m_l can take on only a limited number of values:

$$m_l=0,\ \pm1,\ \pm2,\ \ldots,\ \pm l \tag{7.3.15}$$

That is, the absolute value of m_l can never exceed l.

Plugging in for the value of C_1 removes the unknown constant from the radial equation:

$$-\frac{\hbar^2}{2mr^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)R(r)-\frac{\hbar^2 l(l+1)}{2mr^2}R(r)+V(r)R(r)=E\,R(r) \tag{7.3.16}$$

If the particle is bound, then this equation will give quantized partial wave functions $R_{nl}(r)$ (from the presence of the $l(l+1)$ in the radial equation, it seems clear that the radial wave function will have dependence on the quantum number l as well as the quantum number n that comes from its own differential equation), and we will once again have three quantum numbers: n, l, and m_l.

Just for completeness, we can put the wave function of the stationary state back together:

$$\psi_{nlm_l}(r,\theta,\phi)=AR_{nl}(r)Y_{lm}(\theta,\phi) \tag{7.3.17}$$

All that remains is to plug in the radial potential and use the boundary conditions to solve the differential equation for $R_{nl(r)}$.

While these names may not make sense just yet, these quantum numbers are referred to as the *principle quantum number* (n), the *orbital quantum number* (l), and the *magnetic quantum number* (m_l).

The Free Particle

From our work with the free particle in Cartesian coordinates, it is clear that it is the easiest coordinate system to use when one wishes to discuss plane waves. But as we have mentioned many times before, plane waves are not the only energy eigenstate wave functions. There is in fact a common energy eigenstate for which spherical coordinates are ideal. It is called a *spherical wave*, as it emanates radially outward from (or inward toward) a single point, which of course is the origin of our spherical coordinates. This is more often encountered when discussing light rather than massive particles (as light is frequently emitted from a point source), but as we have seen before, quantum mechanics treats both as "quanta."

For reasons we will see when we discuss the hydrogen atom next, a radially spreading (or contracting) wave function has an l-value equal to zero (and therefore a zero m_l-value as well), and it is left as an exercise to show that this wave function satisfies the free-particle ($V(r)=0$), radially-outward ($l=0$) radial equation:

$$\psi(r)=A\frac{e^{\pm ikr}}{r},\qquad E=\frac{\hbar^2k^2}{2m} \tag{7.3.18}$$

The positive-valued exponent corresponds to the wave that moves radially-outward. Notice that the probability density is simply $\frac{1}{r^2}$, which reflects an inverse-square reduction in intensity when one converts this single particle wave function to a steady-state stream of particles. Also notice that like the plane wave, this wave function cannot be normalized. And finally, as is always the case with energy eigenstates, we can attach the time-dependence by multiplying the stationary-state solution by the usual $e^{-i\omega t}$.

7.4: Angular Momentum

The Dynamical Quantity We Have Ignored

We have spent all of our time in the quantum theory dealing with the three physical quantities of position, momentum, and energy, but there is one other that we studied at length in classical mechanics that has barely been mentioned so far – angular momentum. This is quite an omission, given that it, along with linear momentum and energy, satisfies a very useful conservation principle under the proper conditions. This apparent oversight is even more striking given that Planck's constant, the most fundamental of all physical constants in the realm of quantum mechanics, has units of angular momentum! But avoidance of angular momentum was not an oversight, it just took a lot of background to be able to handle it, and the time is now finally right.

Digression: Types of Vectors

Since Physics 9HA, we have treated angular momentum as though it is a vector. After all, it does have magnitude and direction, right? Well, yes on the magnitude, but we actually defined its direction ourselves with the right-hand-rule – if the majority of humanity had been left-handed, we might have defined it direction as exactly the opposite. It doesn't have a "natural" direction like velocity and momentum does. In fact, angular momentum is an example of a mathematical quantity known as an axial vector, or a pseudovector ("regular" vectors like momentum and velocity are referred to as polar vectors). What distinguishes axial vectors from polar vectors mathematically is how they transform when the coordinate system defining their components is reflected about the origin ($x \rightarrow -x$, $y \rightarrow -y$, $z \rightarrow -z$). Clearly polar vectors will change sign under such a transformation (all of their components flip sign), but see what happens to angular momentum:

$$\left.\begin{array}{l} x \rightarrow -x \\ y \rightarrow -y \\ z \rightarrow -z \end{array}\right\} \rightarrow \vec{L} = \vec{r} \times \vec{p} = (-\vec{r}) \times (-\vec{p}) = +\vec{L} \tag{7.4.1}$$

Building an Angular Momentum Operator

We know from the very origins of our work in quantum mechanics that all physically-observable quantities have quantum-mechanical operators associated with them. We also determined that the operators we needed could be built from a couple of fundamental ones. For example, we built the kinetic energy operator from the momentum operators for each of the three directions:

$$\widehat{KE} \equiv \frac{1}{2m}\left(\hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2 +\right) = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \tag{7.4.2}$$

Well, classically we can express the components of angular momentum in terms of components of position and momentum, so we should be able to do a similar construction from their operators. We have to use some care, however, because whenever we "mix" position and momentum, the uncertainty principle shows up. Let's start with the classical definition:

$$\vec{L} = \vec{r} \times \vec{p} = \left(x\,\hat{i} + y\,\hat{j} + z\,\hat{k}\right) \times \left(p_x\,\hat{i} + p_y\,\hat{j} + p_z\,\hat{k}\right) = (yp_z - zp_y)\,\hat{i} + (zp_x - xp_z)\,\hat{j} + (xp_y - yp_x)\,\hat{z} \tag{7.4.3}$$

Let's focus only on the z-component. If we want to create an operator for L_z, we must replace the classical values with operators:

$$\widehat{L}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x = -i\hbar\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right) \tag{7.4.4}$$

The operators for L_x and L_y come out similarly. Speaking of the operators of the other components of angular momentum, close examination of these reveals a brand-new property that we haven't seen before. back in Section 5.3, we found that the uncertainty principle for two "incompatible" physical quantities is manifested in their operators in that they do not produce the same result when they act on a wave function in different orders. Put another way, physical quantities that cannot simultaneously be measured with arbitrary precision have operators that don't commute with each other. We know that the operators representing the components of the position all commute with each other – we can simultaneously measure the x-position, the y-position, and the z-position without the uncertainty principle making them trip over each other. We can similarly do this with the components of linear momentum, since the order of partial derivatives can be freely swapped. But this is not true of the operators of the components of the angular momentum vector! Let's demonstrate this with the x and y components:

$$\begin{aligned}
\widehat{L}_x\widehat{L}_y\Psi &= \left[-i\hbar\left(y\frac{\partial}{\partial z}-z\frac{\partial}{\partial y}\right)\right]\left[-i\hbar\left(z\frac{\partial}{\partial x}-x\frac{\partial}{\partial z}\right)\right]\Psi = \\
&-\hbar^2\left[y\frac{\partial}{\partial x}\Psi+yz\frac{\partial^2}{\partial x\partial z}\Psi+xz\frac{\partial^2}{\partial y\partial z}\Psi-z^2\frac{\partial^2}{\partial x\partial y}\Psi-xy\frac{\partial^2}{\partial z^2}\Psi\right] \\
\widehat{L}_y\widehat{L}_x\Psi &= \left[-i\hbar\left(z\frac{\partial}{\partial x}-x\frac{\partial}{\partial z}\right)\right]\left[-i\hbar\left(y\frac{\partial}{\partial z}-z\frac{\partial}{\partial y}\right)\right]\Psi = \\
&-\hbar^2\left[x\frac{\partial}{\partial y}\Psi+yz\frac{\partial^2}{\partial x\partial z}\Psi+xz\frac{\partial^2}{\partial y\partial z}\Psi-z^2\frac{\partial^2}{\partial x\partial y}\Psi-xy\frac{\partial^2}{\partial z^2}\Psi\right]
\end{aligned} \tag{7.4.5}$$

We see that the derivatives of the second (leftmost) operator act on the position components present in the second (rightmost) operator. The derivatives make quite a mess in each case, but the point is that the two "messes" are not equal, as can be seen by comparing the two results above – the last 4 terms are all the same, but the first terms differ. If one measures the x-component of angular momentum of a particle precisely, knowledge of the y component of the same angular momentum vector is lost!

Another extremely interesting (and as it happens, useful!) result we get from above is that we can construct an operator of one component of the angular momentum from the other two. Start by subtracting the two results above. This gets rid of the last four common terms and leaves:

$$\left(\widehat{L}_x\widehat{L}_y-\widehat{L}_y\widehat{L}_x\right)\Psi=\hbar^2\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)\Psi \tag{7.4.6}$$

The right side of this equation looks familiar – it is very nearly the operator of the z-component of angular momentum (Equation 7.4.4). Indeed we have:

$$\widehat{L}_x\widehat{L}_y-\widehat{L}_y\widehat{L}_x=i\hbar\widehat{L}_z \tag{7.4.7}$$

There is a shorthand notation frequently used in such situations, called the *commutator* of the operators $\widehat{A}$ and $\widehat{B}$:

$$\left[\widehat{A},\widehat{B}\right]\equiv\widehat{A}\widehat{B}-\widehat{B}\widehat{A} \tag{7.4.8}$$

Thanks to the cyclical nature of the cross-product, we can similarly construct all the components from the other two, so we have:

$$\left[\widehat{L}_x,\widehat{L}_y\right]=i\hbar\widehat{L}_z\,,\quad\left[\widehat{L}_y,\widehat{L}_z\right]=i\hbar\widehat{L}_x\,,\quad\left[\widehat{L}_z,\widehat{L}_x\right]=i\hbar\widehat{L}_y \tag{7.4.9}$$

These are known as the *commutation relations* of the angular momentum component operators.

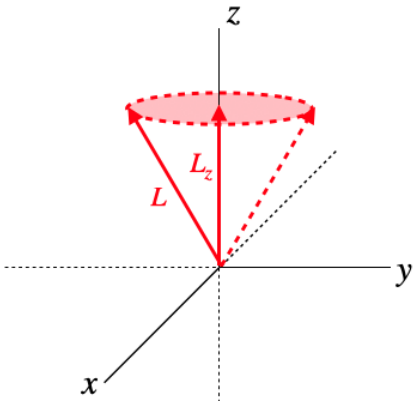
Magnitude of Angular Momentum

We now know how the components of angular momentum relate to each other, but what can be said about the *magnitude* of the angular momentum (or its magnitude-squared, which is more commonly used)? Interestingly, this measurable quantity commutes with all of the components:

$$\left[\widehat{L}^2,\widehat{L}_x\right]=\left[\widehat{L}^2,\widehat{L}_y\right]=\left[\widehat{L}^2,\widehat{L}_z\right]=0\quad\widehat{L}^2=\widehat{L}_x^2+\widehat{L}_y^2+\widehat{L}_z^2 \tag{7.4.10}$$

We will not do so here, but this fact can be proven using purely the commutation relations between the components. The physical meaning is that a particle can be in an eigenstate of total momentum, and one of its components can simultaneously be known to arbitrary precision. But because of the uncertainty relation between the components, the particle can be in an eigenstate of *only one* of those components at a time. We can depict this in a diagram as follows. Let's suppose that we measure both the total magnitude of angular momentum and the z-component precisely, then the uncertain *vector* angular momentum state of the particle looks like the figure below.

Figure 7.4.1 – Uncertain Angular Momentum Vector



Uncertainty in the vector $\vec{L}$ is characterized by the circle shown around the z-axis – every such vector with a tail at the origin and the head landing in the edge of that circle is possible. Note that the *length* of the angular momentum vector is known precisely, as is the z-component, but the x and y components are unknown.

Using Spherical Coordinates

In defining spherical coordinates, we have selected-out the z-direction as special (it is the axis from which the polar angle is measured, and around which the azimuthal angle is measured). By convention, we therefore define that to be the direction for which the single component of angular momentum is known. We can convert the operator $\widehat{L}_z$ into spherical coordinates to get an operator to use in cases where spherical coordinates are in play. The amount of math busywork requires to show this is substantial, but the end result is remarkably simple:

$$\widehat{L}_z = -i\hbar\frac{\partial}{\partial\phi} \tag{7.4.11}$$

If the potential acting on a particle happens to be spherically symmetric, then the separated stationary-state wave function has an azimuthal piece given by Equation 7.3.12, and it's clear that this is an eigenfunction of the z-component of angular momentum:

$$\widehat{L}_z\Psi_{nlm_l}(r,\theta,\phi) = -i\hbar\frac{\partial}{\partial\phi}\left[R_{nl}(r)\,\Theta_{lm_l}(\theta)\,e^{im_l\phi}\right] = m_l\hbar\,\Psi_{nlm_l}(r,\theta,\phi) \tag{7.4.12}$$

The quantum number m_l therefore describes the quantized z component of angular momentum – the eigenvalue is this integer multiplied by $\hbar$.

The operator use in spherical coordinates for $\widehat{L^2}$ is considerably more complicated than the one for $\widehat{L}_z$, and we will not take the discussion this deep in this class. But it is important to know that the polar portion of the separated wave function (the Legendre polynomials, $\Theta_{lm_l}(\theta)$, for which we have also postponed examination to a future, more advanced class) are eigenfunctions of this operator. Specifically:

$$\widehat{L^2}\Psi_{nlm_l}(r,\theta,\phi) = \widehat{L^2}\left[R_{nl}(r)\,\Theta_{l,m_l}(\theta)\,e^{im_l\phi}\right] = l(l+1)\,\hbar^2\,\Psi_{nlm_l}(r,\theta,\phi) \tag{7.4.13}$$

The quantum number l therefore describes the quantized magnitude-squared of angular momentum – the eigenvalue is $l(l+1)$ multiplied by $\hbar^2$. Or, if you like, the eigenvalue for the magnitude of angular momentum is $\left|\vec{L}\right| = \sqrt{l(l+1)}\,\hbar$.

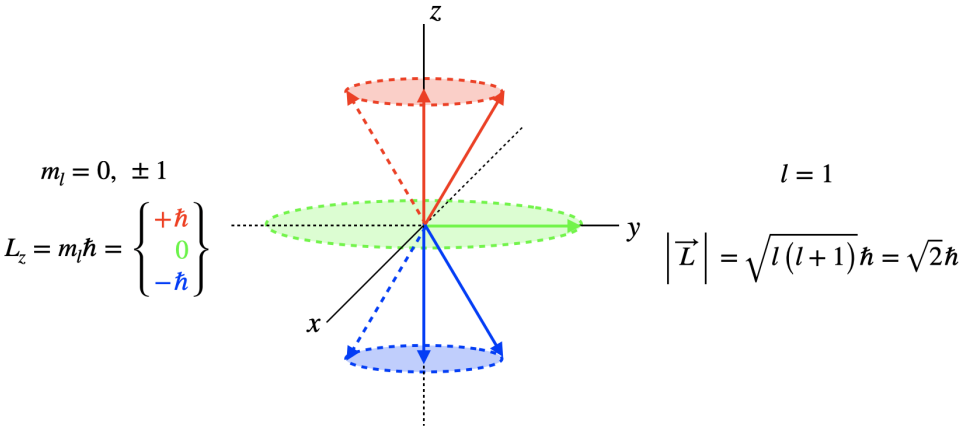
To see further justification for this eigenvalue for the square of the total angular momentum, take a look at the radial Schrödinger equation (Equation 7.3.16). The terms on the left side of the equation that do not include the potential energy must be the kinetic energy (since the right hand side is the total energy). The kinetic energy can be divided into a part that comes from radial motion and a part that comes from tangential motion. We did a similar thing in Physics 9HA when discussing gravitation, and wrote the tangential part of the kinetic energy in terms of the angular momentum in Equation 7.3.8. Comparing these immediately suggests that $L^2 = l(l+1)\,\hbar^2$.

Given that the z-component of the angular momentum must always be *strictly less than* the magnitude of the angular momentum vector (the x and y components can never be zero due to uncertainty), the relation given in Equation 7.3.15, limiting the value of m_l to extremes of $\pm l$, and this magnitude of $\vec{L}$ are perfectly consistent. This leads to the number of possible quantized angular momentum states for a given quantum number l to be:

$$\text{number of quantized angular momentum states} = 2l+1 \tag{7.4.14}$$

For example, for $l = 1$:

Figure 7.4.2 – Quantized Angular Momentum States ($l = 1$)



A Few Brief Words About the Hydrogen Atom

Clearly the radial symmetry of the potential energy of the electron in a hydrogen atom qualifies it for this analysis, but we will not be delving into the details of of the eigenfunctions for this system in this class (rest assured it will be a high priority for a future class in quantum physics!). However, there are some interesting/important facts about the hydrogen atom that are worth sharing.

The electron in a hydrogen atom comes with the three quantum numbers already mentioned: n, l, andm_l. From our study of one-dimensional wells, we are used to having a quantum number related directly to energy eigenvalues. In this case, the energy eigenvalue appears imbedded in the radial wave equation, the eigenfunction of which depends on both n and l. It's not surprising that the magnetic quantum number doesn't effect the energy of the electron, since m_l relates to the quantized z-component of angular momentum, and this depends upon our choice of z axis – the energy levels should not depend upon our choice of a z-axis. But certainly it makes sense that the energy of the electron in the Coulomb potential should depend upon the other two. Strangely, however, it does not! It *only* depends upon the principal quantum number, n. Indeed, Bohr's remarkable result of the energy spectrum of hydrogen with his simple model holds up, even when incorporating the angular momentum properly.

The mathematics (which we didn't cover in detail) of the two angular-momentum quantum numbers gave us a restriction on the possible values of m_l for a given value of l (Equation 7.3.15), and it turns out that still more mathematics (that we also won't cover) results in a restriction on the number of l values for a given n. The values of l are restricted to be only positive, and it turns out that they can never exceed the principle quantum number:

$$l = 0,\ 1,\ 2, \ldots,\ n-1 \tag{7.4.15}$$

When we first started with 3-dimensions, we noted that the extra degrees of freedom can, especially in cases involving symmetry, lead to degeneracies – multiple wave functions responsible for the same particle energy. Clearly this happens for the hydrogen atom. We already know about the $2l+1$ states that have the same l quantum number but different m_l quantum numbers, and now we see that there are multiple orbital quantum numbers for a given principle quantum number. The degeneracy associated with the multiple values of m_l are attributable to the spherical symmetry, but the degeneracy associated with the l-values is a famous example of an accidental degeneracy (recall these are degeneracies that come from coincidence, and not some symmetry (at least not an obvious one).

The degeneracy we get for a given principle quantum number can be counted. There are n possible values of l, and for each of these, there are $2l+1$ values of m_l. This leads to the following number of states for the same energy:

$$\begin{aligned}\text{degeneracy of } n^{th} \text{ energy eigenstate} = &\left[1\ \ (l=0,\ m=0)\right] + \left[3\ \ (l=1,\ m=0,\pm1)\right] + \cdots \\ &+ \left[2l+1\ \ (l=n-1,\ m=0,\pm1,\ldots,\pm l)\right] = n^2\end{aligned} \tag{7.4.16}$$

CHAPTER OVERVIEW

8: Intrinsic Angular Momentum – "Spin"

8.1: Measuring Angular Momentum

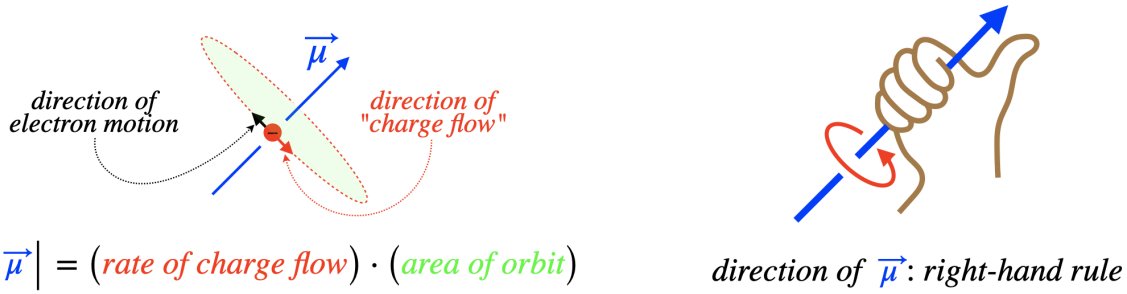
Linking Angular Momentum to Magnetism

While we have some idea now how to handle angular momentum in the quantum realm mathematically, it isn't immediately clear how we are supposed to make measurements of this quantity. For energy it was easy – we looked at emission spectra and used Planck's constant along with the frequency of the light we measured using a diffraction grating. For linear momentum, we looked at the results of collisions. But measuring angular momentum is a tougher nut to crack. We in fact need to look at a different property of particles and derive from what we observe about this property the implications for angular momentum. This other property is called *magnetic dipole moment*, and we will have a quick look at the basics of magnetic forces so that we can understand how this property is used to measure angular momentum.

We have already seen in our discussion of the hydrogen atom that electric charges pull directly toward ot push directly away from each other through the electrical force. Well it turns out that electrical charges exert another type of force (magnetism) on each other, which results from their relative motion. That is, while the mere existence of charge results in an electric field, which in turn results in a force on another particle that possesses charge, more is needed for a magnetic force to occur. The first charge must be moving to produce a magnetic field, and the second charge must be moving to experience a force from that magnetic field. At this point, in light of your Physics 9HB training, alarms are probably going off in your head, and you are wondering, "What does he mean by *moving*?! Relative to what?" Well, I'm afraid the answer to these questions of relativity applied to magnetism will have to wait for Physics 9HD.

We are not even going to look into the details of this magnetic force, as we are interested in one specific case – a charge that is moving in a closed loop while in the presence of a magnetic field. Macroscopically, this could be an electrical current circulating in a conducting loop of wire. Naturally we will be interested in the microscopic version of this – an electron orbiting a nucleus. Rather than looking at the effect of the magnetic field on the electron at any given instant, we simplify things by treating the orbiting electron as a full system (with the location of the electron in its orbit left undefined), and call this system a *magnetic dipole*. The aforementioned magnetic dipole moment is a vector property of such a system. [Interestingly, the idea of treating a magnetic dipole as charge indeterminately smeared-out around its orbit was a classical idea, but actually fits pretty well with what we know of quantum theory!] Here is how the magnetic dipole moment is defined:

Figure 8.1.1 – Magnetic Dipole Moment Definition



The rate of charge flow is the amount of charge (e) divided by the time it takes the charge to complete a full (circular) orbit. This time is the circumference of the orbit divided by the speed of the electron. So the magnitude of the magnetic moment is:

$$|\vec{\mu}| = \left(\frac{e}{\frac{2\pi r}{v}}\right)(\pi r^2) = \frac{evr}{2} \tag{8.1.1}$$

The direction of the magnetic dipole moment is determined by the right-hand-rule applied to the direction of "charge flow". As we are talking about an electron (a negatively-charged particle), the direction of charge flow is opposite to the direction of the electron motion.

Okay, so now we can relate this back to the angular momentum. For a particle of mass m moving in a circle of radius r with speed v, the angular momentum is:

$$\vec{L} = \vec{r} \times \vec{p} \quad \Rightarrow \quad \left|\vec{L}\right| = mvr \tag{8.1.2}$$

We can therefore write the magnetic moment in terms of the angular momentum. Note that the mass of the particle is what matters in the angular momentum, while the charge flow (which is in the opposite direction) is what matters for the magnetic moment. This means that $\vec{\mu}$ and $\vec{L}$ point in opposite directions for the electron, and we have:

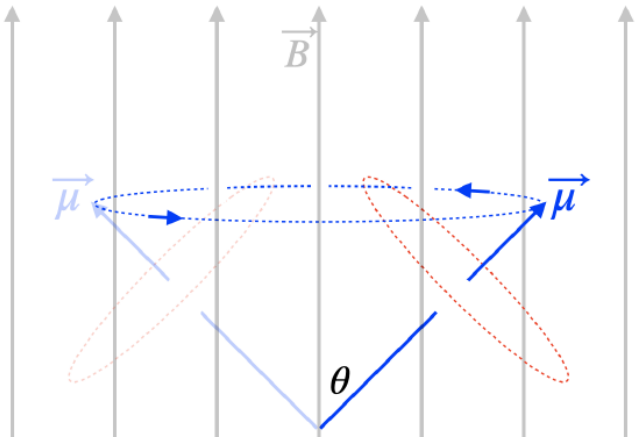
$$\vec{\mu} = -\frac{e}{2m_e}\vec{L} \tag{8.1.3}$$

The magnetic moment and the angular momentum for this particle only differ by a constant, which means that whatever we can do to measure magnetic dipole moment will also tell us the angular momentum.

Precession of Magnetic Dipoles

For our purposes, it will be sufficient to know about two properties of magnetic dipoles. The first of these we have already seen in a different context. In Physics 9HA, we saw what happens when we apply a torque to an object with angular momentum – gyroscopic precession. It turns out that a magnetic field will exert a torque on a magnetic dipole, in the same way that gravity and normal force combined to exert a torque on our gyroscope.

Figure 8.1.2 – Magnetic Dipole Precession (for Electron) Around a Magnetic Field



The torque exerted on a magnetic dipole by a magnetic field $\vec{B}$ turns out to be:

$$\vec{\tau} = \vec{\mu} \times \vec{B} \quad \Rightarrow \quad |\vec{\tau}| = |\mu| \left|\vec{B}\right| \sin\theta \tag{8.1.4}$$

When we studied the gyroscope in Physics 9HA, we determined (Equation 6.2.4) the precession rate in terms of the torque and angular momentum. It's important to keep in mind that in the case of the gyroscope, we had the angular momentum vector pointing perpendicular to the gravitational force. In this case, there is a component of the dipole's angular momentum along the magnetic field. This component will play no role in the precession rate. So defining the $+z$-direction as the direction of the magnetic field, the equation we obtained for gyroscopic precession rate needs to be adjusted so that the angular momentum in the denominator is only the component perpendicular to the z-axis:

$$\Omega = \frac{\tau}{L_{\perp z}} = \frac{\tau}{L\sin\theta} \tag{8.1.5}$$

Plugging in for the torque and the angular momentum gives the precession rate, known as the *Larmor frequency*:

$$\Omega = \frac{\mu B \sin\theta}{\frac{2m_e}{e}\mu\sin\theta} = \frac{eB}{2m_e} \tag{8.1.6}$$

Alert

In the absence of any external constraints, there is no preferred direction in space, and when we seek to describe a spherically-symmetric physical system with spherical coordinates, we are forced to choose an arbitrary z-axis. However, when there is a constraint, such as the applied magnetic field in this case, it is incredibly convenient to define our z-axis in a specific manner. It is universally assumed in quantum theory that the z-direction is chosen to be in the direction of an applied magnetic field.

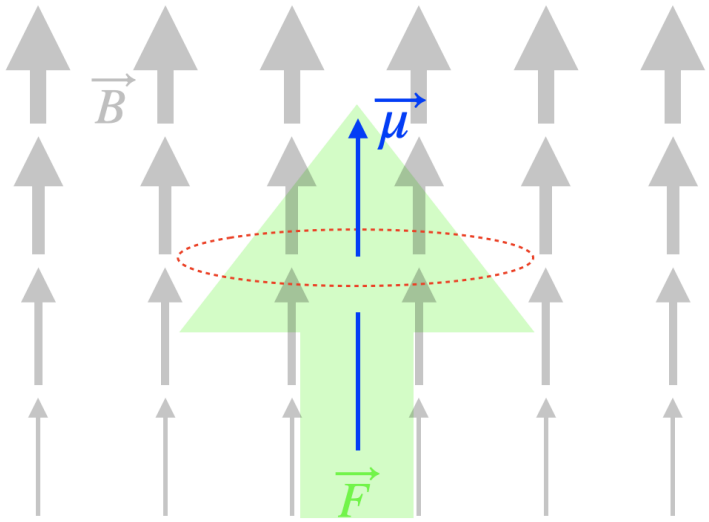
A couple items of note here:

- The Larmor frequency has no dependence on θ, which means it doesn't depend upon the specific angular momentum state of the particle, though the *direction* of the precession does.
- The cone swept-out by the precession bears a striking resemblance to the "uncertainty cone" for angular momentum in Figure 7.4.1. Indeed, while we can measure the rate of precession of the angular momentum vector, we cannot determine its actual direction at any moment in time.
- A comment on the direction of precession indicated in the figure above: The direction of the torque is determined by using the right-hand-rule for $\vec{\mu} \times \vec{B}$, and for the position of the dipole in the diagram, this comes out to be out of the page. But the figure has the dipole precessing *into* the page at that position. The reason is that we use the torque to find the change in the *angular momentum vector*, not the magnetic dipole moment vector. In this case (a negative-charged electron), the dipole moment and angular momentum point in opposite directions, giving the direction of reaction to the torque that is opposite to what it would be for the same diagram depicting a magnetic moment from a positive charge.

Forces on Magnetic Dipoles

While the uniform magnetic field exerts a force at all times on the moving charge, when averaged over a full orbit, there is no net force – only a net torque. But as we have all at some point in our lives stuck a magnet to a refrigerator (or another magnet), we know that it is possible for magnetic fields to exert net forces. The reason the magnetic field above exerts no net force is that it is *uniform*. If we instead introduce a magnetic field that grows (or diminishes) in strength along the direction that it points, then a magnetic dipole pointing parallel to the direction of the magnetic field will experience a net force.

Figure 8.1.3 – Force on a Magnetic Dipole by a Non-Uniform Magnetic Field



The force exerted on the dipole in terms of the gradient of the magnetic field along its direction is:

$$F = \mu \frac{d}{dz} B \tag{8.1.7}$$

Of course, the magnetic dipole moment is not always aligned with the field, so really it is just the z-component of the dipole moment that comes into play. Also, it turns out that it is impossible to have a magnetic field like that described above (you'll have to wait until Physics 9HD to find out why!), where the field is in a single direction everywhere, while getting stronger (or weaker) along that direction. Instead, the direction of the field has to converge or diverge in order for it to get stronger or weaker (respectively) along a given direction. The upshot is that our equation for force on a dipole needs a bit more labeling to be accurate:

$$F = \mu_z \frac{\partial}{\partial z} B_z \tag{8.1.8}$$

As indicated by the diagram, the direction of the force is along the common direction of the field and dipole moment when the field strength is growing in that direction (the derivative is positive).

A Measuring Device

We started this section talking about how to measure angular momentum, and we finally have the means: If we pass a beam of particles through a non-uniform field, then each particle in the beam will be deflected according to its component of magnetic dipole moment along the direction of the field. A device for shooting beams of atoms through a non-uniform field was conceived and constructed in 1922 by Stern and Gerlach. This Stern-Gerlach ("SG") device is as important for demonstrating quantization as the double-slit experiment is for demonstrating wave-particle duality. To see this, consider the results of this experiment...

Randomly-prepared hydrogen atoms have no particular preference for a direction of angular momentum – they can be oriented in any possible direction. So when a beam of these atoms are sent through a non-uniform field, it would seem that each atom should be deflected by whatever force results from that particular atom's z-component of magnetic dipole moment. If its magnetic moment vector happens to be pointing in the direction of increasing field strength, then it will feel a strong force that way. If it is pointing in the direction opposite to the increasing field strength, it will feel a strong force in the opposite direction. If the z-component happens to be perpendicular to the magnetic field, then it will not be deflected at all. And of course, all the z-components between these three extremes should result in a whole continuum of deflections. But this is not what is seen!

We found already that the z-component of angular momentum must be quantized (Equation 7.4.12), and since angular momentum is proportional to magnetic dipole moment, we should *not* see a whole continuum of forces on the particles. Only a few forces are possible – one for each quantum number m_l. It should now be clear why this quantum number is referred-to as the "magnetic quantum number" – it is the one responsible for interactions with magnetic fields.

Similarity to Polaroids

An important lesson learned from these SG devices is what happens when they are used in succession. That is, starting with a randomly-prepared beam of particles, pass it through a SG device, splitting the beam according to the quantized states of angular momentum. Then send *one of the resulting split beams* through a second SG device. What should we expect to happen?

First of all, the original beam was in a mixed-state of z-component of angular momentum, and the SG device had the effect of separating the eigenstates of this operator. Now with all of the beams are eigenstates of that operator, if we send them through a second filter they will not split again, *provided it is the same operator*. Recall that the $\widehat{L_z}$ operator depends upon which way we have defined as our $+z$-direction. If our second SG device is oriented so that its magnetic field points along the same z-axis as the magnetic field in the first SG device, then the second beam will deflect the same as the first did.

If, however, we rotate the z-axis of the second SG device relative to the first one, then the new $\widehat{L_z}$ operator is different from the old one, and the beam that was an eigenstate of the first operator is not in an eigenstate of the new one – the beam splits again! If we pass one of these split beams through a third SG device oriented the same way as the first SG device, we might think that a single beam should emerge – after all, those particles were all eigenstates of the original $\widehat{L_z}$ operator. But this doesn't happen! The particles don't "remember" that they were once in a eigenstate of the original $\widehat{L_z}$ operator – once they are measured using a new z-axis, the emerging beams are eigenstates of *that* $\widehat{L_z}$ operator. As we have seen before, observing the state disturbs the state.

We saw this same sort of behavior for light passing through polaroids, with the exception that the "splitting" of the light results in one of two "beams" being absorbed, while both are passed in the SG device. There is another difference as well, with respect to the fraction of a polarized beam that passes through a second SG device whose orientation differs from the original by some arbitrary angle. We will discuss this difference in the next section.

8.2: It's Not Rotating!

Yet Another Surprising Experimental Result

Armed with what we know about quantized angular momentum, suppose we launch into the following simple experiment: We send a beam of hydrogen atoms in their ground state though a SG apparatus. We know that for hydrogen atoms in their $n = 1$ state, they must have an orbital quantum number $l = n - 1 = 0$, and a magnetic quantum number $m_l = l = 0$. That is, it has no magnetic moment at all, and we expect to see the beam pass undeflected through the magnetic field. Strangely, we don't see this at all!

Maybe we made a mistake somewhere, and we accidentally sent $n = 2,\ l = 1,\ m_l = 0,\ \pm 1$ atoms through? No, this can't be it either, because we don't see three beams emerge (one for each value of m_l – one deflected up, one down, and one undeflected) – we see *two beams* emerge! Okay, things are getting weird. Maybe there are weird complications with the hydrogen atom that we don't understand. Let's remove these complications by taking the proton out of the picture – we'll send just the electrons through the field. To our amazement, we find that there are *still* two beams emerging. We can draw no other conclusion than this: Electrons have a magnetic moment of their own.

Let's Call it "Spin"

It made sense that a hydrogen atom can have magnetic moment – the electron orbits the nucleus, and the radius of this orbit comes into the calculation of what we have defined as magnetic moment in Figure 8.1.1 – the rate of charge flow times the area of the orbit. But an elementary particle like an electron, while it does obviously have the charge needed, *has no radius*, so it is unclear how it can have a magnetic moment.

Okay, so given how little we know about magnetism, maybe there is something we are missing. But there is a more fundamental quantity that we can't forget got us started down this road of using magnetism – angular momentum. If a single point particle has a magnetic dipole moment, it must also possess angular momentum. This is fine when it is orbiting, but our experiment shows that it is somehow intrinsic to the particle. We have actually encountered these two faces of angular momentum before, in 9HA, in Equation 6.1.13. In that case, we found that the total angular momentum of an object breaks down into two terms – one that accounts for the motion of the object through space relative to some reference point (the "orbital" part), and one that accounts for the rotation of the object around its center of mass (the angular momentum "intrinsic" to the object). The problem is that in the case of classical mechanics, even the rotating portion of the angular momentum is at its core orbital, as the particles that make up the rigid object are all "orbiting" the object's center of mass as the object spins, but this is not the case here – the electron has no extension in space, no moment of inertia – *it isn't spinning*!

Despite the conceptual pitfalls of doing so, we nevertheless call this property of the electron (and indeed every elementary particle) *spin*. While it does not have the attributes we normally assume must be present for angular momentum and magnetic dipole moment to exist, electrons do possess these properties. These are distilled down to their very essence, and are fundamentally intrinsic to the particle. An alternative name for spin that is not quite so fraught with faulty intuition is *intrinsic angular momentum*.

Quantization of Spin

Intrinsic angular momentum is perhaps the most purely quantum-mechanical phenomenon we have seen. It really can't be measured in the macroscopic realm like the other quantities we have discussed. When we go to the "classical limit", kinetic energy, momentum, and even orbital angular momentum converge to values we measure easily. But this cannot be done with spin. We can see the *effects* of spin (magnetic materials exhibit magnetic fields at least in part due to spins of a large fraction of particles aligning), but this is really a property we have to look carefully in the microscopic realm to see directly, such as with a SG apparatus.

Spin is a measure of angular momentum, and we already know that angular momentum is quantized. Our SG experiment where we see two beams emerge from a single beam of electrons confirms this. What this experiment also tells us is that the spin angular momentum of electrons comes in only two varieties, with neither of them being zero. We find that no matter what we do to the electrons, we cannot get them to ever split differently in a SG device than into two beams – electrons can only ever have two possible values of the z-component of their intrinsic angular momentum. We generally refer to these two states as *spin up* and *spin down*.

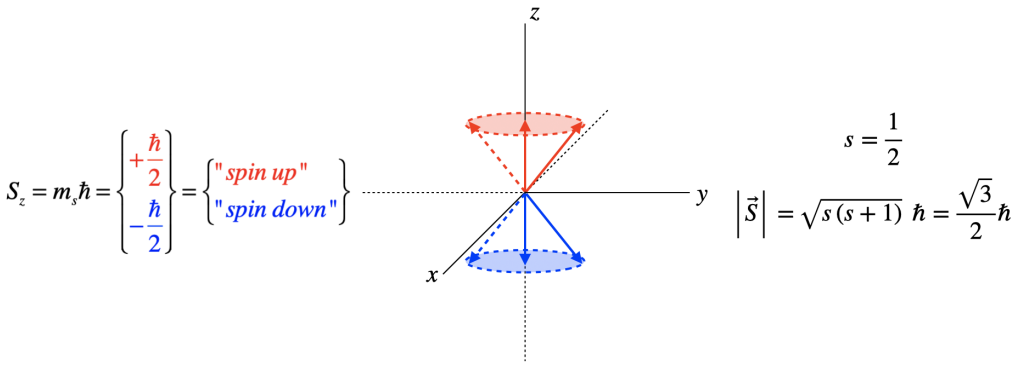
We can do further SG experiments to look into the magnitudes of angular momentum for these two states. For the $l = 1$ orbital case, we found from the deflection of the $m_l = \pm 1$ cases that those z components have a magnitude of $\hbar$. If we compare this to

the results for individual electrons, we find the deflection of individual electrons to be half as great, which means that the z component of angular momentum for an electron is $\pm\frac{\hbar}{2}$. The electron is therefore said to have an intrinsic angular momentum of *spin-½*.

We can now see why we put a subscript "l" on the quantum number "m_l" – the orbital angular momentum is separate from the spin angular momentum. We therefore define "s" to be the quantum number for spin analogous to l for the orbital case, and "m_s" to be the quantum number for the z-component of spin, analogous to the orbital version m_l. As with the orbital case, the possible values for m_s run from $-s$ to $+s$, and change by integer amounts. So for the electron, $s = \frac{1}{2}$, and $m_s = \pm\frac{1}{2}$.

As in the case of orbital angular momentum, the magnitude of the total spin of the electron remains fixed (and the x and y components remain indeterminate), and the resulting picture is similar to Figure 7.4.2:

Figure 8.2.1 – States of Spin-½



Gyromagnetic Ratio

We have come to this description of the intrinsic angular momentum of an electron through the discovery of it having a magnetic moment, so we should revisit the calculation that relates magnetic moment to angular momentum. In our previous calculation that related $\vec{\mu}$ to $\vec{L}$, the radii of the circular orbit divided-out, resulting in Equation 8.1.3. But there was still a remnant of the orbital nature of this calculation – the factor of 2 in the denominator. Given that there is no orbital element to intrinsic spin, it's reasonable to wonder if this factor remains for this case when we replace $\vec{L}$ with $\vec{S}$. The answer is that it does not, and the way this is typically quantified for spin is with a dimensionless constant g, called the *gyromagnetic ratio*:

$$\vec{\mu} = \pm\frac{ge}{2m}\vec{S} \tag{8.2.1}$$

[Note that we have not defined the sign of the charge and left m in the denominator without a subscript, so that this can be applied to any particle, not just electrons.] For electrons, of course the sign is negative, and we find that the value of g is very nearly equal to 2.

Revisiting the Degeneracy of Hydrogen Atom States

Another consequence of electrons having two varieties of spin (up or down) is that the number of degrees of freedom for their wave functions is doubled. We have described an electron in a hydrogen atom with quantum numbers n, l, m_l, but now we see that it doesn't tell us everything – it can either be spin up or spin down, without changing n, l, or m_l. The spin up/down status of the electron does not play a role in the energy levels of a hydrogen atom, so there are now twice as many states as we thought for the same energy level, bringing the degeneracy of the hydrogen atom with principle quantum number n up to $2n^2$.

Fitting Spin into the "Big Picture" of Quantum States

The following question now naturally arises about spin: How does the spin quantum number fit into our previous discussion of the number of quantum numbers matching the degrees of freedom? We already have three quantum numbers for the three spatial dimensions, so where did this additional quantum number come from?

The answer must be that particles have an additional "internal" degree of freedom. Up to now, we have been declaring that the wave function "contains all of the information about the particle," and that we can extract it using operators for physical quantities. But the wave function's information is based on *external* physics – position of the particle in space, potential energy as a function

of position, etc. We can't "derive" the spin property, or come up with an operator that acts on our usual wave function that extracts the information about spin. We need to fundamentally change the wave function itself. We do this by taking-on another component to the wave function, called a *spinor*. This "tacking-on" can be expressed mathematically as follows:

$$\Psi(x,t) \rightarrow \Psi(x,t) \otimes \chi_s \tag{8.2.2}$$

The spinor part of the quantum state "lives" in an entirely different mathematical space than the rest of the wave function (this is emphasized by the "⊗" symbol). Operations like the derivative in the momentum operator have no effect on this part of the wave function.

In the case of an electron (which has only two eigenvalues of spin, "up" and "down"), this part of the state is completely describable with two complex numbers, which are essentially the probability amplitudes of measuring each of these two states. As was the case for the spatial part of the wave function, a general spin state is a linear combination of the eigenstates, and since there are only two of these for the electron, we have:

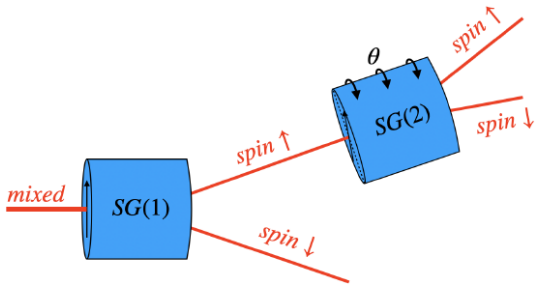
$$\chi_s = \alpha\,\chi_\uparrow + \beta\,\chi_\downarrow\,, \quad |\alpha|^2 + |\beta|^2 = 1 \tag{8.2.3}$$

The complex-valued numbers α and β are just the spin equivalents of the "C_n's" we used to expand the spatial wave function in terms of its eigenstates. Naturally the value $|\alpha|^2$ is the probability that a SG device will measure the electron's spin to be "up", and $|\beta|^2$ is the probability that a SG device will measure the electron's spin to be "down".

Successive Spin Measurements

In the previous section, we noted that using a SG device as a "spin filter" bears similarities to our discussion of passing light through polaroids. In the latter case, we derived a relation called Malus's law, which gives the intensity of light that emerges from a polaroid after incoming polarized light passes through it, in terms of the angle formed between the incoming light polarization and the orientation of the polaroid. We know that particles that are eigenstates of spin "up" through one SG device are not in an eigenstate of a second SG device that does not have its magnetic field pointing in the same direction, so there must be a relation that is equivalent to Malus's law for spin-½. Equivalent, yes, but not identical, as it turns out...

Figure 8.2.2 – Successive SG Filters at a Relative Angle



If the orientation of the second SG device's magnetic field differs from that of the first's by an angle θ, and the spin "up" beam that comes from the first SG device is passed through the second, then what started as an eigenstate of spin becomes a mixed state (purely spin "up" becomes a linear combination of spin "up" and spin "down"), with the mixing coefficients (choosing real-values only) being:

$$\chi_s\,[\text{after } SG(2)] = \alpha\,\chi_\uparrow + \beta\,\chi_\downarrow = \cos\left(\frac{\theta}{2}\right)\chi_\uparrow + \sin\left(\frac{\theta}{2}\right)\chi_\downarrow \tag{8.2.4}$$

The difference with Malus's law becomes clear when we create a "polaroid" out of an SG device by simply absorbing all beams that come out spin "down", so that it only permits a fraction of the particles to pass through, as a polaroid does. When you flip $SG(2)$ over, the beam that was previously measured as spin "up" is now measured as spin "down", which means that as a "polaroid" turning it 180 degrees blocks all of the particles. In the case of light through a polaroid, a 180 degree rotation allows all of the light to pass.

The probability of a single spin-½ particle getting through $SG(2)$ as spin "up" after passing through $SG(1)$ as spin "up" when the fields of $SG(1)$ and $SG(2)$ are rotated relative to each other by an angle θ is:

$$P(\uparrow) = |\alpha|^2 = \cos^2\left(\frac{\theta}{2}\right) \tag{8.2.5}$$

The half-angle argument of the cosine function differs from the full-angle result for Malus's law.

8.3: Fermions and Bosons

Not All Spins Are "One-Half"

The only particle we have discussed with regard to spin is the electron, but we can talk about the intrinsic spin of any particle. Some particles are spinless ($s = 0$), others have half-integer spins ($s = 1/2,\ 3/2$), and still others have non-zero integer spins ($s = 1,\ 2$). Note that this latter group (which includes the photon, $s = 1$) has an integer *intrinsic* angular momentum, and this should not be confused with the integer-valued *orbital* angular momentum we have seen. That is, while the only possible orbital quantum numbers are integers, spin quantum numbers can be either integer or half-integer.

Two Categories of Particles

We will see shortly that particles with half-integer spin behave quite differently from those with integer spin, and this different behavior results in a specific "division of labor" in the universe. The exploration of the nature of forces in quantum theory has resulted in a model that describes forces as exchanges of particles. For example, we know that accelerated electrons emit light (photons). These accelerations result from electromagnetic forces on the electrons. This "exchange particle" model ties these together neatly, by explaining that the mechanism of force itself comes about due to photon exchanges between charged particles. The mathematics of this theory is much too involved to go into in this class, but the concept interests us for the following reason: It turns out that all the particles that mediate force (e.g. the photon for the EM force, and the graviton for the gravity force) have integer spin, while all the particles that comprise matter (which we think of as the stuff that feels the force) have half-integer spin (electrons, protons, and neutrons are all spin-½).

This stark division of the particle types, the tasks they perform, and as we will see, the properties they exhibit, inspires us to give these two categories specific names (named after physicists who studied them in the earliest years of quantum theory). Half-integer spin particles are known generically as *fermions* (named for Enrico Fermi), and integer spin particles are called *bosons* (named for Satyendra Nath Bose).

Identical Particles

When we first discussed the notion of a "quantum state" of a particle, we said that it is a collection of all the information available about that particle, and that this information can be organized into a wave function (with a spinor attached). Suppose now that we consider two particles in tandem. These now constitute a new system, for which we can define a new quantum state. The puzzle we need to solve is how to express this new state in terms of the states of the individual particles. We are not unfamiliar with this idea, thanks to our experience with statistical/thermal physics in 9HB. In kinetic theory, we describe the motions of the individual particles, but then as we step back and look at the collection as a whole, we define a "thermodynamic state". In that case, we bridge the gap between the collective state and individual states with Maxwell-Boltzmann statistics, which relates things like the distribution of energy of individual particles to the temperature of the collection as a while.

> **Alert**
>
> *Throughout our discussion of multi-particle quantum states, we will assume (much like we do with ideal gases) that the particles involved do not interact with each other (such as through the coulomb potential). The particles can be individually affected by some external potential, but the effects we will be discussing come only from the fact that they are combined into a single system.*

In our quantum mechanics version of linking multiple particles to individual ones, we have a few new problems. For one, we ultimately can only deal with probabilities. This was true for the collection in the Maxwell-Boltzmann case as well, but with quantum theory this probabilistic nature goes all the way down to individual particles. In thermodynamics, we assume that if we could track all of the particles, we can precisely predict the evolution of the system, and we only use probability because we cannot practically accomplish this particle-tracking. In quantum mechanics, we can't track the individual particles *even in principle*.

There is an even more important issue that comes about due to out inability to track particles in quantum theory. We assert that *all* of the information about a particle's quantum state is contained in its wave function. Wave functions track quantities like energy and angular momentum with their quantum numbers, but there is no quantum number that acts as a particle identifier. If two particles of the same type (say electrons) are commingling in a confined region (say in an infinite square well), their wave functions are overlapping, and when we measure the position or energy of one of the particles, we cannot tell which of the two particles we have looked at. In classical physics the particles don't have labels either, but the particles are distinguishable because we can just

watch the particles closely and know which one we are measuring the energy of. This property of like particles being indistinguishable in quantum mechanics will turn out to have profound consequences.

So how do we treat a quantum state of multiple particles mathematically? Let's start by considering a simple case – two *distinguishable* particles, like an electron and a proton (remember, we are still assuming that these particles are not interacting, so perhaps a neutron is a better choice for the other particle). Each of these particles has their own wave function, but we want to form a single wave function, which stores all the information about this "system". The answer lies in how we incorporated spinors into our states. In that case, we asserted that the spinor "lives" in another space, and the intrinsic space of the spin is subject to operators that don't act on the "extrinsic" wave function. This combination of entities is known mathematically as a *direct product space*, and can also be applied to separate particles:

$$\Psi\left(\text{two-particle system}\right) = \Psi\left(\text{particle } A\right) \otimes \Psi\left(\text{particle } B\right) \tag{8.3.1}$$

The real key here is the probabilities. The probability density comes out to be just a product of probability densities:

$$P\left(\text{two-particle system}\right) = \left|\Psi\left(\text{two-particle system}\right)\right|^2 = \left|\Psi\left(\text{particle } A\right)\right|^2 \otimes \left|\Psi\left(\text{particle } B\right)\right|^2 \tag{8.3.2}$$

We can now ask for the probability of particle A being found between x_1 and $x_1 + dx_1$ *and at the same time* finding particle B between x_2 and $x_2 + dx_2$:

$$P\left(A \text{ near } x_1 \text{ and } B \text{ near } x_2\right) = \left|\Psi_A\left(x_1\right)\right|^2 dx_1 \left|\Psi_B\left(x_2\right)\right|^2 dx_2 \tag{8.3.3}$$

Here we have dropped the direct product notation, because the variables x_1 and x_2 keep things straight. We want to express this as a single probability density of a system, and clearly there are now two inputs (x_1 and x_2), so this looks like:

$$\left|\Psi_{system}\left(x_1, x_2\right)\right|^2 dx_1 dx_2 = \left|\Psi_A\left(x_1\right)\right|^2 dx_1 \left|\Psi_B\left(x_2\right)\right|^2 dx_2 \tag{8.3.4}$$

Okay, so this is all simple enough. It makes sense, because the two particles are not interacting, so the two events (measuring one particle at one position and the other at another) are independent, and the probability of two independent events occurring is just the product of the probabilities of each one occurring.

But now if the two particles in question are identical, we can't say "particle A near x_1 and particle B near x_2" anymore. We can only say, "one of the particles near x_1 and the other particle near x_2". The independence necessary for the simple multiplication of probabilities is lost, because in the first case, finding A near x_1 and B near x_2 is different from finding B near x_1 and A near x_2, but when the particles are indistinguishable, these events are the same.

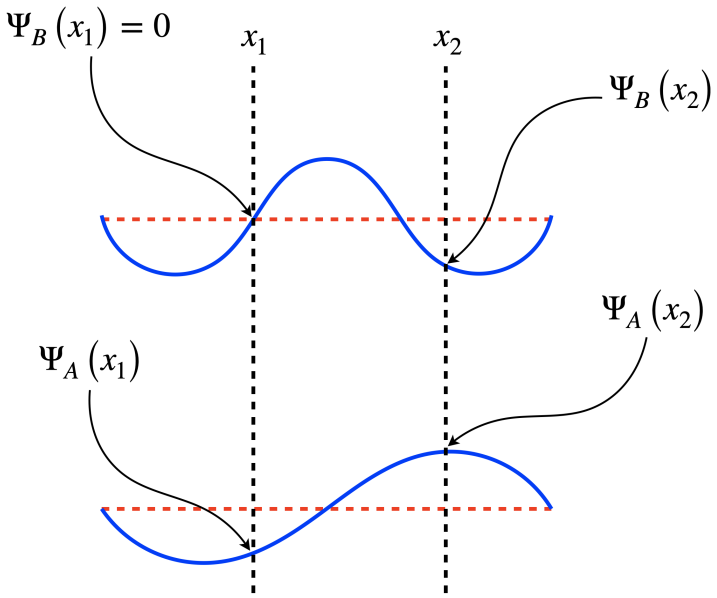
Exchange Symmetry

With the particles indistinguishable, the probabilities measured by the quantum state should not change if we swap the positions of the two particles. That is, if we swap the variables x_1 and x_2 in the two-particle wave function, the probability density should not change:

$$\left|\Psi_{system}\left(x_1, x_2\right)\right|^2 dx_1 dx_2 = \left|\Psi_{system}\left(x_2, x_1\right)\right|^2 dx_1 dx_2 \tag{8.3.5}$$

It should be clear that this is not the case for Equation 8.3.4, but in case it is not, consider the diagram below, which depicts two partial wave functions for a one-dimensional two-particle system, and two positions along the x-axis. You can think of particle A as being in the $n = 2$ eigenstate of an infinite square well, and particle B being in the $n = 3$ eigenstate. We don't actually know *which* particle is in each of these states, of course, and that's the point.

Figure 8.3.1 – Partial Wave Functions of a Two Particle State



If we plug in the values $\Psi_A(x_1)$ and $\Psi_B(x_2)$ into Equation 8.3.4, we get some non-zero probability density. But now suppose we swap the positions of the two identical particles – they are indistinguishable, so it should not change the probability density. But $\Psi_B(x_1)$ is *zero*, which means that the probabilities don't match when the particles switch positions.

So how do we fix the construction of the two-particle wave function from the partial wave functions? We do this by not linking the each specific particle with a particular position. In the original solution with separation of variables, we can acknowledge that although there are two quantum numbers (one for each degree of freedom, the degrees of freedom in this case coming from two particles being present), we have to leave it undetermined which one goes with each wave function. It's possible to show that there are two ways to construct the two-particle wave function from the partial wave functions, that both satisfies the stationary-state Schrödinger equation and yields the same probability when the positions are swapped. They are these two states:

$$\Psi_+(x_1, x_2) = \frac{1}{\sqrt{2}}[\Psi_A(x_1)\Psi_B(x_2) + \Psi_A(x_2)\Psi_B(x_1)] \tag{8.3.6}$$

$$\Psi_-(x_1, x_2) = \frac{1}{\sqrt{2}}[\Psi_A(x_1)\psi_B(x_2) - \Psi_A(x_2)\Psi_B(x_1)] \tag{8.3.7}$$

The subscripts "+" and "−" stand for "symmetric" and "antisymmetric," respectively, for obvious reasons. These two cases provide the two possibilities for *exchange symmetry* for quantum particles. The constant in front of these is there to normalize the two-particle wave function assuming that the partial wave functions are already normalized.

Pauli Exclusion Principle

So now we have asserted that there are two types of particles in the universe (fermions with half-integer spin, and bosons with integer spin), and that when it comes to two-particle states, the individual wave functions can be combined in either symmetric or antisymmetric fashion to make a system wave function. There is no reason to believe that these two elements of quantum theory should be linked in any way, and yet remarkably they are! Proving this is far from straightforward (so we will not do it in this class), but the link is this: If our system of two identical particles happens to consist of bosons, then their wave functions observe symmetric exchange symmetry, and if the two identical particles are fermions, then their wave functions observe antisymmetric exchange symmetry.

The most striking consequence of this fact becomes clear when we consider two identical fermions with the same quantum numbers. For example, suppose we have two electrons in the ground state of a one-dimensional infinite well, both with spin up. These particles have identical wave functions, and when we combine them with an antisymmetric exchange, we just get *zero* for the system's wave function. A zero wave function means zero probability, so this means that we can never witness such a thing. When we put two electrons into this box, and we measure the energies and spins of the particles, we will never find that the quantum numbers are all the same for both. This *exclusion principle* is attributed to Wolfgang Pauli, and has many far-reaching consequences.

8.4: Statistics of Identical Particles

Multiple Particles in the Same Potential Well

The increased degeneracy introduced by the additional degree of freedom that comes from spin addresses the number of unique states for the same energy level of a single, but where the addition of this additional quantum number really gets interesting is when we put several particles together into a common potential. Again, we continue to assume that the particles do not interact with each other.

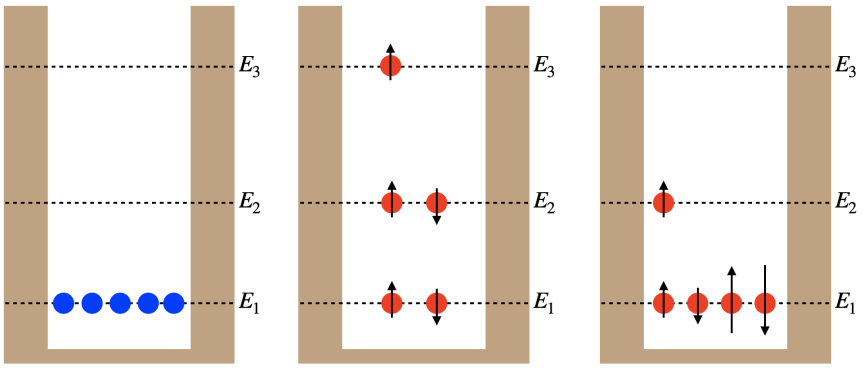
Let's take as an example a one-dimensional infinite square well (we aren't very creative, this is the first example we use for everything). Let's place five identical particles into this "box," such that they reach an energy eigenstate. [Keep in mind we are talking about a multi-particle quantum state here.] The question we want to answer is, what is the ground state energy of this configuration?

It turns out that we can't answer this without knowing what types of particles we are putting into the box. We know from our separation of variables work (and frankly, just from energy conservation), that the total energy of the multi-particle state will be the sum of the energies of the individual particles, so we might think that the lowest total energy occurs when each particle is in its individual ground state. But not so fast!

What if the particles are electrons? These are spin-$\frac{1}{2}$ fermions, which means that there is zero probability of having two of them with the same quantum numbers. There are two quantum numbers present here – one for the single dimension, and one for the intrinsic spin degree of freedom. So we can "fit" two electrons into their lowest individual energy states if they have opposite spins, but we can't get the third electron into the ground state without violating the exclusion principle, so this third electron must reside in its first excited state. We can also get a fourth electron into the first excited state, but then the fifth must be in the second excited state.

If the particles are bosons, then no problem, all three particles can reside in the ground state at once, and that will be the lowest total energy state for the system. Have we solved this problem for all particles? No! Recall that not all fermions are spin-$\frac{1}{2}$; higher spin quantum numbers are also possible. If we use spin-$\frac{3}{2}$ fermions, then there are *four* different spin states available: $m_s = \pm\frac{1}{2}, \pm\frac{3}{2}$. With four different states available for a single energy level, there is plenty of room to fit all of these fermions in their individual ground states.

Figure 8.4.1 – Multi-Particle Ground State According to Particle Spin



Revising Boltzmann – The Statistics of Identical Particles

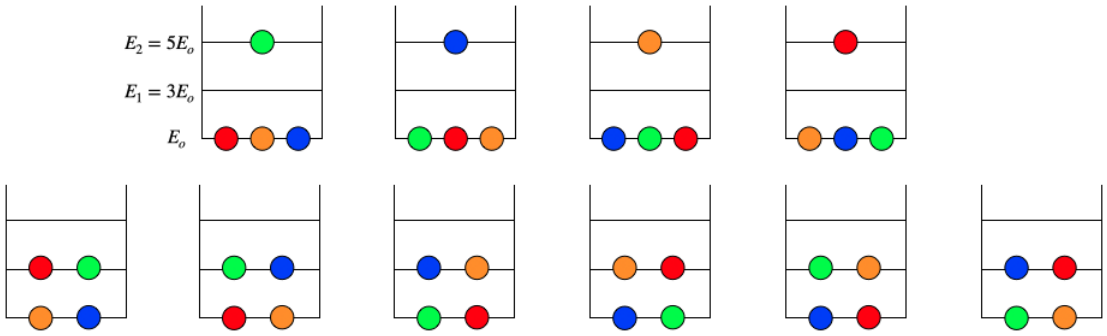
The quantum mechanics of identical particles has a profound effect on statistical physics. In 9HB, we encountered the Boltzmann distribution and applied it to computing entropy from multiplicities of microstates. The whole subject of the Boltzmann distribution boiled down to this: Give a collection of particles a certain total amount amount of energy. Each particle gets its own share of that energy, and those "shares" are distributed through a wide range of values. We can compute the probability (when the collection reaches equilibrium) that any given particle will get an amount of energy in a given infinitesimal range, by using the fact that the distribution will be at equilibrium when the maximum multiplicity is reached.

But here is the rub: The Boltzmann distribution assumes that all of the particles are distinguishable. In the non-quantum world in which Boltzmann operated, the motions of all particles could in principle be tracked. We can see why things change by considering 4 particles, this time all in a one-dimensional harmonic oscillator potential. We wish to compute the multiplicity of the state where the system has a total energy of $8E_o$, where of course $E_o = \hbar\omega_c = \hbar\sqrt{\frac{\kappa}{m}}$. There are only two ways that the energies of these particles can add up to $8E_o$: Three particles in the ground state, and one particle in the second-excited state, or two particles in each the ground state and first excited state:

$$\begin{aligned} 3\cdot E_o + 1\cdot E_2 &= 3\cdot \tfrac{1}{2}\hbar\omega_c + 1\cdot\left(2+\tfrac{1}{2}\right)\hbar\omega_c = 4\hbar\omega_c = 8E_o \\ 2\cdot E_o + 2\cdot E_1 &= 2\cdot \tfrac{1}{2}\hbar\omega_c + 2\cdot\left(1+\tfrac{1}{2}\right)\hbar\omega_c = 4\hbar\omega_c = 8E_o \end{aligned} \tag{8.4.1}$$

Let's start by counting the multiplicity that these available states provides for distinguishable particles (Boltzmann statistics). We can place 4 distinguishable particles into energy levels such that their total energy adds up to $8E_o$ in ten different ways:

Figure 8.4.2 – Combinations of Distinguishable Particles



If we wanted to know (say) the probability of selecting a particle at random in the ground state, we would divide the number of particles in the ground state by the total number of possibilities, and find it equals 0.6. When we increase this to a very large number of particles, we get the Boltzmann probability of finding a particle at energy E_n for a collection of particles at temperature T:

$$P(E_n) \propto \frac{1}{e^{E_n/k_BT}} \tag{8.4.2}$$

Now let's remove the condition that the particles are distinguishable. We'll look at the case for bosons first. When we make all the "particles" in our diagram the same color, we get only two different set-ups:

Figure 8.4.3 – Combinations of Indistinguishable Bosons

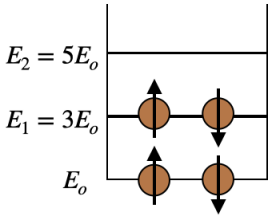
Here we see that the probability of randomly selecting a particle in the ground state is different from the case with distinguishable particles – it is 0.625. We won't go into the combinatorics to prove it here, but for a large number of particles, the difference between the probabilities for indistinguishable bosons (called *Bose-Einstein statistics*) and the distinguishable particle case comes out to be an extra "-1" term in the denominator:

$$P_{BE}(E_n) \propto \frac{1}{e^{E_n/k_BT} - 1} \tag{8.4.3}$$

[*Note: This and the case to follow have been slightly over-simplified to emphasize the effect of the change of statistics in the denominator, with the removal of a number called the chemical potential. In the case of a "photon gas" (which satisfies Bose-Einstein statistics, as photons are spin-1), this chemical potential is zero, and this expression is accurate.*]

Okay, what about the case of spin-½ fermions? In this case, we know that no more than two can occupy the same energy state, thanks to the exclusion principle. This gives exactly one configuration for our 4-particle case:

Figure 8.4.4 – Combinations of Indistinguishable Spin-½ Fermions



Now the probability of randomly selecting a particle in the ground state is different from the previous two cases – it is now 0.5. For a large number of particles, the difference between the probabilities for indistinguishable fermions (called *Fermi-Dirac statistics*) and the distinguishable particle case comes out to be an extra "+1" term in the denominator:

$$P_{FD}\left(E_n\right) \propto \frac{1}{e^{E_n/k_BT}+1} \tag{8.4.4}$$

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: UCD: Physics 9HC – Introduction to Waves, Physical Optics, and Quantum Theory

Webpages: 54

All licenses found:

- CC BY-SA 4.0: 81.5% (44 pages)
- Undeclared: 18.5% (10 pages)

By Page