

UCD: PHYSICS 9D - MODERN PHYSICS



Tom Weideman
University of California, Davis

University of California, Davis
UCD: Physics 9D - Modern Physics

Tom Weideman

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, 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.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 06/12/2026

TABLE OF CONTENTS

Licensing

1: Sound

- 1.1: Fundamentals of Sound
- 1.2: Doppler Effect
- 1.3: Interference Effects

2: Foundations of Special Relativity

- 2.1: The Relativity Principle
- 2.2: The Nature of Time
- 2.3: More Thought Experiments
- 2.4: Paradoxes

3: Kinematics in Special Relativity

- 3.1: Spacetime Diagrams
- 3.2: Lorentz Transformation
- 3.3: Velocity Addition
- 3.4: Electricity and Magnetism

4: Dynamics in Special Relativity

- 4.1: Momentum Conservation
- 4.2: Energy Conservation

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

- 5.1: Blackbody Radiation
- 5.2: The Photoelectric Effect
- 5.3: Compton Effect
- 5.4: Double-Slit Experiment
- 5.5: Matter Has Wave Properties, Too!
- 5.6: Summarizing this Wave/Particle Mess

6: The Universe is Inherently Probabilistic

- 6.1: Basics of Probability Theory
- 6.2: Continuous Probability Distributions and Probability Density
- 6.3: The Uncertainty of Random Outcomes
- 6.4: Physical Measurements with Random Outcomes
- 6.5: Incompatible Measurements

7: Matter Waves

- 7.1: The Schrödinger Wave Equation
- 7.2: States of Definite Energy
- 7.3: Operators and Observables
- 7.4: Eigenstates and Eigenvalues

8: One-Dimensional Models

- [8.1: Particle-in-a-Box, Part 1](#)
- [8.2: Particle-in-a-Box, Part 2](#)
- [8.3: The Finite Square Well](#)
- [8.4: Tunneling](#)
- [8.5: The Quantum Harmonic Oscillator](#)
- [8.6: The Bohr Model of the Hydrogen Atom](#)

9: Quantum Theory in Three Dimensions

- [9.1: Schrödinger's Equation in 3-Dimensions](#)
- [9.2: 3-Dimensional Models](#)
- [9.3: Central Forces](#)
- [9.4: Angular Momentum](#)

10: Intrinsic Angular Momentum – "Spin"

- [10.1: Measuring Angular Momentum](#)
- [10.2: It's Not Rotating!](#)
- [10.3: Fermions and Bosons](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

CHAPTER OVERVIEW

1: Sound

[1.1: Fundamentals of Sound](#)

[1.2: Doppler Effect](#)

[1.3: Interference Effects](#)

This page titled [1: Sound](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

1.1: Fundamentals of Sound

Why Sound?

Physics 9D is a class about modern physics, so why is the first chapter in the textbook dedicated to the phenomenon of sound? Three reasons:

1. There is typically a long period of time that elapses between taking Physics 9B (where general wave phenomena are first studied) and taking Physics 9D, where wave physics is used extensively. By returning to the specific wave phenomenon of sound, you are given an opportunity to review some of the general features of waves, while applying them to the specific physical conditions present for sound.
2. Studying sound provides a useful historical context for modern physics. Sound was a wave phenomenon that was studied extensively prior to the modern era of physics, and this knowledge was both a help and a hindrance to physicists trying to unravel the mysteries that came around in the transition period between the 19th and 20th centuries. We can benefit from following their journey.
3. Examining properties of sound provides a useful contrast to the unusual aspects of waves that arise in relativity and quantum physics.

As Physics 9D progresses, the reader is encouraged to sort out which characteristics of sound waves are closely paralleled in modern physics, and which are vastly different (but may nevertheless be useful as an analogous model).

Sound Waves in Air

Sound can travel through any phase of matter – solid, liquid, or gas. Like other mechanical waves, it depends upon a restoring mechanism that returns particles in the medium to equilibrium after they are displaced. But unlike a wave in a string, for which the restoring mechanism is perpendicular to the direction of wave motion, the direction of sound's restoring mechanism is parallel to the direction of the wave motion – sound waves are longitudinal rather than transverse. As we [discussed in Physics 9B](#), this comes about as a result of compressions and rarefactions. In one-dimension, this can be visualized as compressing and expanding coils of a slinky, but sound waves can travel in three dimensions, which makes the picture slightly more challenging to grasp.

We will primarily focus our exploration on sound waves in air, mainly because that is the way that we usually encounter them. What is not commonly noted is that sound waves in air are fundamentally different from sound waves in liquids and solids. As we noted in Physics 9B, gases are collections of particles that, to a very good approximation, don't interact with each other. If one particle in a medium doesn't interact with a neighboring particle, it seems strange that a wave can propagate through that medium at all. After all, waves require some sort of "restoring force" that returns the medium to its equilibrium state after it is displaced away from it. So how does the gas medium oscillate as a wave goes by, if the particles are not experiencing forces to make them oscillate?

The answer is "probability and statistics." Whenever there is an imbalance in the populations of particles in a gas (as there is when there is a low-density rarefaction next to a high-density compression), on average more randomly-moving particles enter the low-density region than leave it, and more particles leave the high-density region than enter it. So the low-density region naturally (through sheer probability) grows in density, while the high-density region drops in density, providing a statistical rather than mechanical "restoring force." Note that for solids (and to a lesser extent, liquids), the particles *do* interact with each other, and the restoring forces are exactly that – forces, but not so for sound through gases. As you can see below, the sound wave of compressions and rarefactions propagates along, even as the particles of the gas fly around randomly.

Figure 1.1.1 – Sound Wave in Air



Alert

Frequently textbooks and other sources, in their discussion of sound in air, refer to the oscillatory displacement of particles in the medium. It is okay to refer to the "average displacement" of a particle from the equilibrium point, but individual particles in a gas are flying all over the place, not actually bouncing back-and-forth. The distinction between these can easily be lost, leading to a lot of confusion.

For these reasons, when we express a wave function for sound in air, it will have units of either density or pressure. Note that when we detect a sound with our ears, it is the variations in pressure in the medium that causes our eardrums to vibrate.

Sound Wave Properties

The speed of sound in air at standard temperature and pressure is around 343m/s . For air and other fluids, the sound wave velocity dependence on the medium is very similar to that which we found for a transverse wave on a string. The density changes from a linear density to a volume density (which we denote with a ρ), and the tension is replaced by a constant known as **bulk modulus**. The velocity relation looks like:

$$v_{\text{sound in fluid}} = \sqrt{\frac{B}{\rho}} \quad (1.1.1)$$

Sound will also travel through a solid, but in that case the interactions of the particles are different than in a fluid, and the constant that takes the place of tension is a different one: **Young's modulus**. But the formula looks the same:

$$v_{\text{sound in solid}} = \sqrt{\frac{Y}{\rho}} \quad (1.1.2)$$

We will not explore the exact nature of the bulk and Young's moduli – simply knowing that they play the same role for fluids and solids respectively as the tension plays for a transverse wave on a string will suffice for our purposes.

Alert

At a very young age, children in science classes learn that sound travels faster through water than through air, and faster through solids than through water. This often leads to the erroneous conclusion that sound travels faster in media that are more dense. Indeed, the opposite is true, and in fact it is the greater bulk or Young's modulus that accounts for the faster speed of sound.

The intensity of a sound wave also obeys the rule-of-thumb for intensity – the intensity is proportional to the square of the amplitude. Specifically, it turns out that for an amplitude measured in pressure, the intensity is given by:

$$I = \frac{A^2}{2\sqrt{\rho B}} \quad (1.1.3)$$

Alert

It is important to note that the amplitude of pressure is the difference between the maximum pressure of the compression (or the minimum pressure of the rarefaction) and the ambient (equilibrium) pressure.

As with any other wave, the dimensions into which a sound wave spreads also determines how the intensity varies with distance from the source. That is, a sound that expands spherically outward has its intensity dissipate according to the inverse-square law. This explains why sound made into a closed tube (like those that can be found in playgrounds for children to play with) will remain so much louder despite the distance the sound travels – the sound is not allowed to spread out spherically, and is instead reflected back into the direction of the tube. Even shouting through a cone or cupped hands has some effect in this regard.

The Decibel Scale

The human ear is very sensitive to detecting sound. How loud a sound is depends upon the amplitude of the vibration of the eardrum, which is determined by how much energy the sound wave transfers to the eardrum per second. This of course depends upon the intensity (which is multiplied by the area of the eardrum to get the power transferred), and it turns out that the range of intensities that the ear can detect before it starts becoming painful is quite large. A healthy human ear can hear sounds with intensities as low as 10^{-12}W/m^2 (known as the **threshold of hearing**), and starts to feel pain around 1W/m^2 . A range of 12 orders of

magnitude is quite large, making it more convenient to count the powers of ten rather than the exact values. A logarithmic scale has therefore been devised that works as follows:

We start with a benchmark value – the threshold of hearing – which will translate to a zero value in the logarithmic scale (so the power of ten will be zero). Then just convert every intensity to a ratio with this benchmark, and take the logarithm (to base 10):

$$\beta = (10 \text{ dB}) \log_{10} \left(\frac{I}{I_o} \right) \quad (1.1.4)$$

The number yielded by just the logarithm of the ratio is described as the number of "bels" of the loudness of the sound. It is traditional to multiply this by number by 10, so that the unit describing the loudness is **decibels**. Note that the threshold of hearing results in zero decibels, while the pain threshold occurs at 120 dB.

Alert

Sometimes the decibel level of sound is referred to as the sound's "intensity." Strictly speaking this is not accurate, but as there is a one-to-one correspondence between an intensity and a decibel level, it doesn't cause problems, especially if the context of this use of the term "intensity" includes some mention of a number of decibels.

Example 1.1.1

A medieval village has a bell located in a tower in its central square which is rung to warn the townspeople of emergencies, such as raiding parties from nearby regions. If the loudness of the bell heard by villagers in the town 500 ft (about 1/10 mile) from the tower is 20 dB, then about how far from the tower does the sound carry (i.e. at what distance does the bell become barely audible)? Assume that there is negligible energy dissipated from the sound wave due to obstacles and the atmosphere.

Solution

The decibel scale is logarithmic, which means that every time the decibel level changes by 10 dB, the intensity changes by a factor of 10. The bell can barely be heard at the threshold of hearing, which is 0 dB, which means that the decibel level can afford to drop by 20 dB, and the intensity can drop by two factors of 10 (i.e. drop by a factor of 100). The sound from the bell expands outward spherically, so the intensity drops off according to the inverse-square law. Therefore to drop by a factor of 100, the distance must increase by a factor of 10. So the bell can barely be heard at a distance of one mile.

Example 1.1.2

A speaker at the north end of a round football stadium emits a sound at a single frequency. A listener in the center of the stadium hears the sound at an decibel level of 35 dB. Speakers in phase with the north end speaker are then turned on at the east, south, and west ends of the stadium, with all four speakers emitting sound at the same power output. Find the decibel level of the sound heard at the center of the stadium from all four speakers combined. Assume no thermal dissipation of sound wave energy into the air.

Solution

The sound waves coming from the four speakers all start in phase at the same time, and travel the same distance, so when they reach the common point at the center of the stadium, they are in phase, and interfere constructively. [Note that the direction of motion of the sound is irrelevant, as the contributions to the density of the air is what is superposing.] With four identical waves in phase, the superposed wave will have four times the amplitude of each individual wave. Multiplying the amplitude of the sound wave by 4 results in an intensity that is increased by a factor of 16. Now all we need to do is determine how much of a change this means for the decibel level (which is not a factor of 16!):

$$\begin{aligned} \beta_{\text{four speakers}} &= (10 \text{ dB}) \log \left(\frac{I_{\text{four speakers}}}{I_o} \right) \\ &= (10 \text{ dB}) \log \left(\frac{16 I_{\text{one speaker}}}{I_o} \right) \\ &= (10 \text{ dB}) \log 16 + (10 \text{ dB}) \log \left(\frac{I_{\text{one speaker}}}{I_o} \right) \\ &= 12 \text{ dB} + 35 \text{ dB} = 47 \text{ dB} \end{aligned}$$

This page titled [1.1: Fundamentals of Sound](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

1.2: Doppler Effect

Sound Source in Motion

All waves begin at some source, and have an effect some distance away at something we will call a "receiver." Sound waves in particular exhibit this source/receiver relationship very well: Something (a "source") vibrates, varying the air pressure in its vicinity in some fashion. This pressure variance then propagates as a sound wave to another place, where the varying pressure causes another object (the "receiver") to vibrate. We have already seen how the separation distance and the inverse-square law relates the vibration of the source to the vibration of the receiver in terms of their energies.

Let's consider the following three properties of a sound wave generated at a source and detected at a receiver:

- period (or frequency)

The period of the wave according to the source is simply the time span between the generation of each wave crest, and the frequency measured by the source is the inverse of this number. As each compression reaches the receiver, a compression occurs, so the time that elapses between crests "washing over" the receiver is the period of the wave as measured by the receiver, and the inverse of this period is the frequency measured by that receiver.

- wavelength

The spatial separation of the wave crests is the wavelength of the sound wave. This is not a measurement that "belongs" to either the source or receiver – a simple snapshot of the particles in the air can be used to measure this distance.

- wave velocity

The two values above – the period and wavelength – are related to each other according to the speed of the wave. That is, the source calculates the speed of the wave to be the ratio of the wavelength and the period measured by the source. The receiver calculates the wave speed as the ratio of the wavelength and the period that it measures.

The question we seek to answer here is, "How does motion of the source through the medium affect these quantities, if at all?" To answer this question, we'll start with the simple case where there is no movement through the medium, and create a model that simplifies comparisons of what happens at the sender and receiver. We'll assume that the sender is point source (depicted with a red dot in the figure below), and that the receiver is a point in space as well (depicted with a blue dot in the figure below). The radiating circles in the figure represent sound wave crests (regions of maximum density or pressure), propagating outward from the source. Every time a new crest is created at the source, the red dot flashes, and every time a wave crest is detected, the blue dot flashes. The time between red dot flashes is the period of the wave according to the source, and the time between blue dot flashes is the period of the wave according to the receiver.

Figure 1.2.1 – Source and Receiver Stationary



The source and receiver have been placed exactly three wavelengths apart, which has the effect of synchronizing their flashes, which makes it clear that the time between red dot flashes is the same as the time between blue dot flashes. Of course, the time interval between flashes would be the same even if the source and receiver were separated by a different distance and the flashes were not simultaneous. The point is that the source and receiver agree upon the period (frequency) of this wave. There is no reason that the source and receiver should ever disagree about the wavelength of the wave, as a snapshot can be taken at any moment and a meter stick can be used to measure this quantity. Also it is clear that the wave moves away from the source at the same speed that it moves toward the receiver. So both source and receiver agree upon measurements of all three of the quantities in the equation $v = \lambda f$.

Now let's suppose that the source is moving through the air toward the stationary receiver as the sound is emitted. After a wave crest leaves the source, it continues spherically outward from the point where the source was when it emitted the sound, but the source moves before it emits the next wave crest, which results in the spherical wave crests not being concentric. The result is depicted in the figure below.

Figure 1.2.2 – Source Moving Toward Stationary Receiver



While the source is emitting a crest at the same regular time intervals as in the stationary case above (the time between red dot flashes is the same), the wave crests that reach the receiver are closer together. The speed of the wave crests according to the receiver is the same as before (the only difference is that they start at different places), so the time interval between blue dot flashes is shorter than before. It is clear that the blue dot is flashing twice as frequently as the red dot. So the source and receiver do *not* agree upon the period (frequency) of this wave!

The wavelength in this case is tricky – clearly it depends upon where you measure it. In front of the source it is much shorter than behind it. For the purposes of this discussion, just the wavelength along the line joining the source and receiver matters, and as before a snapshot can be taken, and the distance measured between wave crests is well-defined. So both the source and receiver agree upon the wavelength, but disagree upon the frequency, so what happens to the relation $v = \lambda f$? Clearly the source and receiver cannot agree on the speed of the wave. Of course they don't! The wave moves at a fixed velocity *through the medium*, and the receiver measures this speed because it is not moving through the medium. But the source *is* moving through the medium, which means it will measure a slower speed for the wave crests. Indeed, the source sees the crests moving away from it (in the direction specified) slower than the receiver sees the same crest coming toward it. So the source measures a slower wave speed and lower frequency for the wave than the receiver, and both measure the same wavelength, which allows both to satisfy the equation $v = \lambda f$.

Ultimately we would like a relationship between the frequencies measured by the source and receiver. To this end, let's make the following definitions:

f_s	$\equiv$	the frequency of the wave measured by the source. Equals the inverse of the period of the wave, T_s
f_r	$\equiv$	the frequency of the wave measured by the receiver. Equals the inverse of the period of the wave, T_r
v_s	$\equiv$	the velocity of the source relative to the medium
v	$\equiv$	the velocity of the sound within the medium

We can compute the wavelength in front of the source using these quantities. A wave crest is emitted, and in the time it takes for a second one to be emitted (T_s), it travels a distance of vT_s . At this point, a second wave crest is emitted, but it is not emitted at the same position as the previous one. This one is emitted from a position that is closer to the receiver by the amount that the source has moved in the same time period: $v_s T_s$. The wavelength is the distance between these two crests:

$$\lambda = vT_s - v_s T_s = (v - v_s) T_s = (v - v_s) \frac{1}{f_s} \quad (1.2.1)$$

Now we can use the wavelength (which is the same for the receiver as the source) to relate the two frequencies:

$$\lambda = \frac{v}{f_r} \Rightarrow \frac{v}{f_r} = (v - v_s) \frac{1}{f_s} \Rightarrow f_r = \left(\frac{v}{v - v_s} \right) f_s \quad (1.2.2)$$

The fraction is greater than one, so this formula agrees with our observation that the frequency of the wave measured by the receiver is greater than the frequency sent by the source.

Suppose the source was moving away from the receiver. Then the wavelength is *increased*, which means that the time span between wave crests reaching the receiver is increased, and the receiver measures a lower frequency than the source. The amount that the wavelength is expanded is found the same way that the amount it was reduced in the previous case, and the effect is that the two terms in Equation 2.2.1 are added rather than subtracted. This changes the

final answer such that the minus sign in the denominator becomes a plus sign. We can therefore summarize the relationship between the source and receiver frequencies (known as the **doppler effect**) for motion along a line as:

$$f_r = \left(\frac{v}{v \mp v_s} \right) f_s \quad (1.2.3)$$

The upper (−) sign refers to the source moving toward the receiver, and the lower (+) sign refers to it moving away from the receiver. If the source is not moving directly toward or away from the receiver, then things get only slightly more complicated. Note that the wavelength is different for every angle the velocity of the source makes with the line between the source and receiver. While this is not a particularly difficult extension to the doppler formula, it's more than we will need for our purposes, and we will examine it no further.

Example 1.2.1

A pedestrian standing on a corner hears the siren of a police car coming directly toward her. At this point in time, the car is 700 m away and she hears a decibel level of 25 dB. The policeman in the car hears the siren at a frequency of 1000 Hz, while the pedestrian hears it at a frequency of 1100 Hz. Find the decibel level the pedestrian hears 22 s later. Assume the police car maintains a constant speed and that sound exits the siren radially into 3-dimensions. The speed of sound in air is $344 \frac{m}{s}$.

Solution

The loudness of the siren will of course increase as it gets closer to her. We can use the inverse-square law for intensity to determine how much louder it is when it is closer 22 seconds later, but to determine how much closer it is, we first have to figure out how fast it is going. We do this using the doppler effect. We are given the source's frequency and the listener's frequency, so we use those to compute the velocity of the source, noting that the listener is not moving, and the source is moving toward the listener:

$$f_r = \left(\frac{v}{v \mp v_s} \right) f_s \Rightarrow v_s = v \left(1 - \frac{f_s}{f_r} \right) = \left(344 \frac{m}{s} \right) \left(1 - \frac{1000 Hz}{1100 Hz} \right) = \frac{344}{11} \frac{m}{s}$$

The distance traveled by the police car during the 22 seconds is therefore:

$$d = vt = \left(\frac{344}{11} \frac{m}{s} \right) (22s) = 688m$$

Subtracting this from the original distance tells us that the car is now a mere 12 m away. We use this fact to determine the change in the intensity of the sound from the inverse-square law (Equation 1.3.14):

$$I_1 r_1^2 = I_2 r_2^2 \Rightarrow I_{close} = \frac{r_{far}^2}{r_{close}^2} I_{far} = \frac{(700m)^2}{(12m)^2} I_{far} = 3400 I_{far}$$

Now we need to express this in the logarithmic scale for decibels. It is simplest to find the change in decibel level:

$$\beta_{close} - \beta_{far} = (10dB) \log \left(\frac{I_{close}}{I_o} \right) - (10dB) \log \left(\frac{I_{far}}{I_o} \right) = (10dB) \log \left(\frac{I_{close}}{I_{far}} \right) = (10dB) \log 3400 = 35dB \Rightarrow \beta_{close} = 25dB + 35dB = 60dB$$

Sound Receiver in Motion

Okay, now that we know what effect a moving source has on the frequency measured by a stationary receiver, we'll look at the opposite scenario – the effect on the frequency when the receiver is moving toward and away from the source. The overall effect is similar, in that moving toward the source increases the frequency and moving away from the source decreases the frequency, but the analysis is slightly different. With the source stationary, the wave crests are not squeezed closer together or stretched farther apart. The wavelength is simply determined from the source's frequency and the speed of sound in the medium:

$$\lambda = \frac{v}{f_s} \quad (1.2.4)$$

If the receiver is moving into the crests (toward the source), then the crests are moving toward the receiver at a relative speed of $v + v_r$, where now v_r is the velocity of the receiver. The time it takes between crests hitting the receiver (blue light flashes) is the distance traveled (one wavelength) divided by the relative speed, so:

$$T_r = \frac{\lambda}{v + v_r} \quad (1.2.5)$$

Putting these two equations together gives us the relationship between the frequencies:

$$\frac{1}{f_r} = T_r = \frac{\lambda}{v + v_r} \Rightarrow f_r = \left(\frac{v + v_r}{v} \right) f_s \quad (1.2.6)$$

The fraction is greater than one, so the frequency measured by the receiver is indeed higher than that of the source. As before, if the receiver is moving away, then there is a sign change from this case, giving:

$$f_r = \left(\frac{v \pm v_r}{v} \right) f_s \quad (1.2.7)$$

As before the upper sign indicates motion toward, and the lower sign motion away.

Example 1.2.2

Sound is emitted from a stationary source, and is detected by a stationary receiver. Naturally both measure the same frequency. The source now starts moving away from the receiver, and the frequency of the sound heard by the receiver is shifted lower by 25%. If the receiver had instead moved away from the source at the same speed, by what percentage would the frequency of the sound shift down?

Solution

We start by determining the speed we are talking about here. From the source-moving formula, if the receiver frequency is 25% lower, then it is $\frac{3}{4}$ of the source frequency:

$$f_r = \left(\frac{v}{v + v_s} \right) f_s \Rightarrow \frac{v}{v + v_s} = \frac{3}{4} \Rightarrow v_s = \frac{1}{3}v$$

So the source is moving away at a speed of one third the speed of sound. Now we only need to compute the change in frequency when the receiver moves away from the stationary source at the same speed (i.e. plug in $\frac{1}{3}$ for v_r):

$$f_r = \left(\frac{v - v_r}{v} \right) f_s \Rightarrow \left(\frac{v - \frac{1}{3}v}{v} \right) = \frac{2}{3}$$

So if the source moves away at the same speed, then the frequency drops by one third (33%).

Combinations of Motions

The next natural question that arises is, "What if both the source and receiver are moving?" In this case, we can actually just "stack" these two results. When we calculated the effect of the receiver moving toward the source, we started with the wavelength of the sound. In that particular case, the wavelength resulted from a stationary source, but if it hadn't, the derivation would have been the same. So if the wavelength is compressed or stretched by the motion of the source, we use that wavelength, and get the answer from there. That is, instead of plugging Equation 2.2.4 (which expresses a stationary source) into Equation 2.2.5, we use Equation 2.2.1 (which expresses a moving source) instead:

$$\left. \begin{aligned} \lambda &= (v \mp v_s) \frac{1}{f_s} \\ \frac{1}{f_r} &= \frac{\lambda}{v \pm v_r} \end{aligned} \right\} \Rightarrow f_r = \frac{v \pm v_r}{v \mp v_s} f_s \quad (1.2.8)$$

Once again the top sign handles motion toward the other, and the bottom sign handles motion away. There are a couple of checks we can do on this result. First, putting in $v_r = 0$ or $v_s = 0$ gives the same equations we found above for a stationary receiver and source, respectively. Second, we notice that if both objects are moving in the same direction at the same speed through the air, then one is moving toward while the other is moving away, and these speeds are equal, so the numerator equals the denominator, and both measure the same frequency.

Another interesting combination that comes up often is the *echo*. If a source is moving toward (say) a stationary wall, and the sound sent by the source is reflected off the wall and heard by the source (which now has become a receiver), how does the emitted frequency compare with the received frequency? The important concept to understand here is that when a wave strikes a new medium (in this case, sound going from air into a solid wall), the property of the wave that is maintained is the frequency. This is because if we measure the time between successive wave crests hitting the new medium, the same time elapses between wave crests emitted by that new medium. This applies to both reflection and transmission. So whatever frequency of sound is received by the wall then becomes the frequency of the sound transmitted by the wall in the echo.

So the sound received by the original sender after an echo undergoes two successive doppler effects. The wave that strikes the wall is a different frequency from the what was sent by the moving source. Then upon reflection, that new frequency is transmitted (with the wall treated as a new stationary source) and measured by the original sender (which is now a moving receiver).

Example 1.2.3

An automated flying drone comes equipped with an ultrasonic sensing device. This device emits sound pulses with a frequency of 100 kHz to probe its surroundings by detecting echoes of those pulses from nearby objects. The drone flies along the x -axis in the $+x$ direction, when it detects an echo from a UFO that is directly in front of it that is also moving along the x -axis. The onboard computer for the drone immediately logs the following data:

airspeed of the drone:	$24.0 \frac{m}{s}$
frequency of the echoed sound pulse:	$109 kHz$

- Are the drone and the UFO moving toward or away from each other as this data is being recorded? Explain.
- Find the direction in which the UFO is moving through the air ($+x$ or $-x$).
- Find the speed of the UFO through the air.

Solution

a. The frequency of the wave is increased in the course of the round-trip. We know that if the drone and UFO were moving at the same speed in the same direction, then the UFO would "hear" the same frequency as the drone emits, would then echo back that same frequency, and again since they are moving the same speed and direction, the drone would "hear" the same frequency echoed by the UFO, which means there would be no doppler shift at all. Clearly then if the UFO slows down (or reverses direction) or the drone speeds up so that they are moving toward each other, the effect will be to doppler shift the detected echo to a higher frequency than the emitted sound pulse. So the drone and UFO are moving toward each other.

b. There are a couple ways to do this. We will look at one way here, and the second way will be given in the answer to part (c). Suppose the UFO is stationary. The doppler effect for the echoed sound would then be found in the usual 2-step manner: UFO hears a doppler-shifted sound, reflects that frequency back, and the drone hears that sound doppler-shifted again. The first shift is due to a moving source, and the second shift is due to a moving listener, so:

$$\left. \begin{aligned} f_{\text{echo}} &= \left(\frac{v}{v - v_{\text{drone}}} \right) f_{\text{emitted}} && \text{[moving drone sends sound]} \\ f_{\text{received}} &= \left(\frac{v + v_{\text{drone}}}{v} \right) f_{\text{echo}} && \text{[moving drone receives echoed sound]} \end{aligned} \right\} \Rightarrow f_{\text{received}} = \left(\frac{v + v_{\text{drone}}}{v - v_{\text{drone}}} \right) f_{\text{emitted}} = \left(\frac{344 \frac{\text{m}}{\text{s}} + 24 \frac{\text{m}}{\text{s}}}{344 \frac{\text{m}}{\text{s}} - 24 \frac{\text{m}}{\text{s}}} \right) (100 \text{kHz}) = 115 \text{Hz}$$

So if the UFO were stationary, then the frequency shift would be 15 kHz, and this is more than the 9 kHz. For the frequency shift to be reduced from the case of when the UFO is stationary, the UFO must be moving away from the drone, which is in the $+x$ direction.

c. We do this in two steps. The first incorporates the doppler shifted sound heard by the UFO, then for the echo that frequency becomes the source, and it is doppler shifted again when it is heard by the drone's detector. Unlike part (b), the doppler shifts are due to motion of both the source and the listener each time. Since we know the UFO and drone are getting closer, we'll assume that the UFO is moving in the $-x$ direction (toward the drone):

$$\left. \begin{aligned} f_{\text{echo}} &= \left(\frac{v + v_{\text{ufo}}}{v - v_{\text{drone}}} \right) f_{\text{emitted}} \\ f_{\text{received}} &= \left(\frac{v + v_{\text{drone}}}{v - v_{\text{ufo}}} \right) f_{\text{echo}} \end{aligned} \right\} \Rightarrow f_{\text{received}} = \left(\frac{v + v_{\text{drone}}}{v - v_{\text{drone}}} \right) \left(\frac{v + v_{\text{ufo}}}{v - v_{\text{ufo}}} \right) f_{\text{emitted}}$$

We are given the emitted and received frequencies, as well as the drone speed, so we can solve for the UFO speed. The result of the algebra is:

$$v_{\text{ufo}} = -0.0268 v = -9.2 \frac{\text{m}}{\text{s}} \quad (1.2.9)$$

Notice that the sign came out negative. At the beginning we assumed that the UFO was moving toward the drone (in the $-x$ direction). The negative sign indicates that the UFO is in fact moving in the opposite direction, in agreement with what we determined in part (b).

This page titled 1.2: Doppler Effect is shared under a CC BY-SA license and was authored, remixed, and/or curated by Tom Weideman.

1.3: Interference Effects

Standing Waves

Sound can create a variety of interference effects, like any other wave. Among those interference effects are standing waves. These are formed in precisely the way described for strings in [Physics 9B](#), with two traveling waves reflecting back-and-forth between two endpoints. In the case of sound in a gas, it isn't immediately clear what constitutes "fixed" and "free" endpoints, since sound waves in gas do not involve particles of media that can ever be held in place. We have to broaden our notion of what it means to have a fixed boundary to mean simply that whatever physical quantity plays the role of "displacement" for a wave is unchanging. In the case of sound, this would mean pressure or density. We will see in a moment how this can be.

We are used to talking about one-dimensional standing waves, so it is fair to ask how this can be set up for sound in a gas. If a region of the gas has a pressure higher than the ambient pressure, the gas will naturally expand into the lower-pressure region, so to create a 1-dimensional standing sound wave, we must set up circumstances so that the compressed regions doesn't expand into all three dimensions. We can do this by producing the sound within a hollow pipe. The fixed or free boundary conditions at the ends of the pipe then depend upon whether the end of the pipe is open or closed, but which case is fixed, and which is free?

Let's look first at a closed end. When a compression propagates its way toward a closed end, the means for the compression to restore itself to equilibrium is restricted – it can't continue forward. The compression grows *even greater* than the amplitude of the wave, just as the displacement of a string at the point of reflection from a free end grows above the amplitude (see [Figure 1.5.5](#) in the 9B textbook). So in fact a closed end of a hollow pipe represents a "free" end for a standing sound wave.

The open end is quite different. In fact, it does not behave like an end at all, but rather like a transition point. We know this is true if we follow the energy – a sound wave headed for an open end will transmit energy out of that open end, as the compressions and rarefactions are transmitted to the region of the gas outside the pipe. So why would there be any reflection back into the pipe at all? As in the case of going from slow-to-fast or fast-to-slow medium, the mathematics of the boundary interaction is beyond this course, but the result is the same. The obvious objection here is that the speed of the wave within the air in the pipe is no different from the speed of sound outside the pipe. This is true, and in fact it requires a bit of a revision to that observation. Perhaps a better way of describing the reflection/transmission phenomenon is to say that there is at least partial reflection whenever *the wave equation governing the wave changes*. One way for the wave equation to change is for the wave velocity to change. Another is for the wave to go from being confined to one dimension to three dimensions. This is exactly what happens when the sound in a pipe emerges from an open end.

We know that waves in three-dimensions spread the energy very fast, causing the amplitude to drop-off in proportion to the distance. It is therefore not a bad approximation to claim that at the open end (or slightly beyond it) is at a fixed pressure/density (the ambient pressure/density). This means that if we are forced to choose, the open end of the pipe to a good approximation behaves like a fixed end, and the sound wave that reflects back into the pipe is phase-shifted.

From these considerations, we now know that standing sound waves in a pipe create pressure antinodes at closed ends, and pressure nodes at open ends. Armed with this information, we can use all the same machinery regarding standing waves that we learned in [Physics 9B](#). What is interesting about the case of sound is how this can be used to create tones in organ pipes and wind instruments. These devices make use of two special aspects of standing waves in pipes. The first we have already mentioned – an open end allows for some transmission of sound waves, which means that we can hear the tone produced without having to crawl inside the pipe. The second has to do with the idea of *resonance*.

Digression: Resonance

Resonance is a very important concept in many fields of physics, and we unfortunately don't have time to cover it in great detail here, though it will come up again in both Physics 9C and 9D. The main idea is that if vibrating systems interact with each other, the amount of energy that can be transmitted from one to the other is largely dependent upon how closely the "natural frequencies" (think of masses on springs with frequencies that look like $\omega = \sqrt{\frac{k}{m}}$) of those systems match. If the frequencies are a good match, then the superposition of the displacements of the two systems leads to constructive interference. If they don't match, then the displacements don't synchronize, and there is very little overall constructive interference. A good analogy is pushing a child on a swing. If you give them a push forward every time they reach the peak of their backswing, then the frequency of your pushes matches the natural frequency of the swing, and energy is transmitted to the swing. If, however, you were to push with a different frequency, then some of the pushes would add energy to the swing, but many others would push the swing forward as it is coming backward, taking energy away from the swing.

If we get air moving near the end of a pipe, like when we blow into a flute, the result is turbulent flow. This adds energy to the system, but the waves created come in a wide variety of frequencies. All of these sound waves travel the length of the pipe, partially reflecting and transmitting when they get there. But as we have seen in our study of standing waves, only those waves that have one of the harmonic wavelengths will exhibit the constructive interference required for a standing wave. Those waves that do have the right frequency exhibit resonance, building energy for the standing wave at that frequency. The result is that of the many sound waves that escape the pipe, those at a resonant frequency of the pipe (determined by its length) have by far the most energy, and are the only sounds heard. It turns out that by far most of the energy goes to the fundamental harmonic, though overtones can also often be heard. To modulate the frequency of the sound that escapes therefore becomes a matter of changing the length of the pipe. An organ designates a pipe for every key on the keyboard; a slide trombone allows the player to physically expand the length of the pipe; valves and holes in other wind instruments also serve this same purpose. Note also that all of these instruments rely upon the turbulent flow to provide the spectrum of sound waves, whether it is through a vibrating reed, vibrating lips, or air forced across an open end. (Note: Air forced **into** an open end does not induce much turbulence compared to air forced **across** the opening.)

Lastly, it should be noted that to produce a sustained tone, one must continually add energy. This is because energy is always exiting the pipe via the transmitted wave. The rate at which energy is added equals the rate of energy escaping via sound waves, while the energy in the standing wave within the pipe remains constant.

Example 1.3.1

A string is plucked above a pipe that is open at one end, and the fundamental harmonic tone is heard coming from the pipe. If the closed end of the pipe is now opened, how must the tension of the string be change in order to excite the new fundamental harmonic?

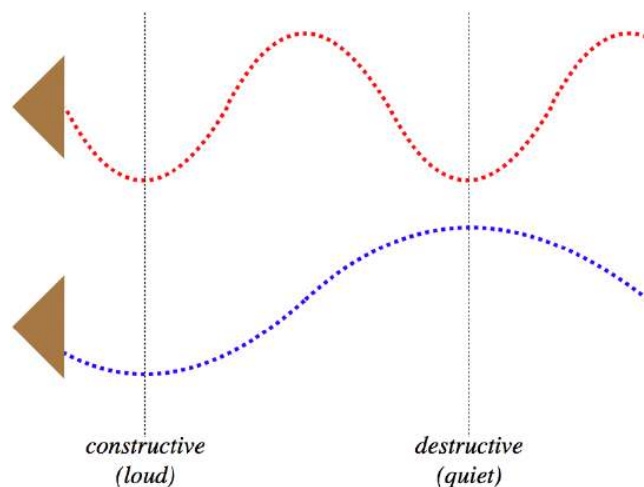
Solution

Open ends of pipes act as fixed points (nodes) for standing sound waves. The fundamental harmonic of the pipe with one end closed fits one quarter of a wave between the ends of the pipe (node to first antinode), while the first harmonic with both ends open fits a half-wavelength (node to node). Therefore opening a closed end shrinks the wavelength of the fundamental harmonic by a factor of 2. The speed of the sound wave is unchanged, so its fundamental harmonic frequency rises a factor of 2. The standing sound wave is being driven by (i.e. is getting energy from) the standing wave in the string, so their frequencies must match for this resonance to occur, which means that when the end of the pipe is opened, the string's standing wave frequency must also rise by a factor of 2. The string has not changed length, so the only other way to change its frequency is to change the speed of the traveling wave on the string. To double the speed of the string wave (and therefore double the frequency), the tension must be quadrupled, since the wave speed on the string is proportional to the square root of the tension, and the linear density of the string cannot be changed.

Beats

Up to now, all of our cases of interference involve waves that have identical frequencies. But a very interesting phenomenon emerges when two sound waves interfere that have slightly different frequencies (actually any two different frequencies will do, but the phenomenon is easier to detect when the frequencies are within 1 or 2 hertz, for reasons we will see). We'll start with another over-simplified diagram like those we have used above. This time, the two waves emerging from the speakers will have different frequencies, and therefore different wavelengths (though for simplicity, we will assume they have the same amplitude).

Figure 1.3.4 – Superposing Two Sound Waves of Different Frequencies



The diagram is of course a snapshot at a moment in time. If, at this moment in time, a listener is positioned at the left vertical line, then the superposition of the two waves results in constructive interference, which means that the sound heard is loud. If, on the other hand, someone is positioned at the right vertical line, then destructive interference results in silence. But now suppose that one remains at the right vertical line for a short time after this snapshot is taken. Both waves are traveling at the same speed (they are both in the same medium), so the rarefactions that coincide at the left vertical line will soon be at the right vertical line. That is, a person listening at the right vertical line will, at one moment, hear silence, and then a short time later, a loud tone. This pattern will in fact repeat itself with regularity, and this pulsing of the sound is referred to as *beats*.

The math that governs this phenomenon results from a straightforward application of trigonometric identities. We start with a wave function for each wave. As we are talking about sound for which the "displacement" is pressure, we will represent the wave function with the variable $P(x, t)$. Note we are also keeping things simple by staying in one dimension:

$$P_1(x, t) = P_o \cos\left(\frac{2\pi}{\lambda_1}x - \frac{2\pi}{T_1}t + \phi_1\right), \quad P_2(x, t) = P_o \cos\left(\frac{2\pi}{\lambda_2}x - \frac{2\pi}{T_2}t + \phi_2\right) \quad (1.3.1)$$

We are interested in what is happening to the sound at a single point in space (i.e. we want to show it gets it gets louder and softer), and any point will do, so for simplicity we'll choose the origin, $x = 0$. This reduces our two functions to:

$$P_1(t) = P_o \cos\left(-\frac{2\pi}{T_1}t + \phi_1\right), \quad P_2(t) = P_o \cos\left(-\frac{2\pi}{T_2}t + \phi_2\right) \quad (1.3.2)$$

These functions are oscillating with different frequencies, so they are out-of-sync. We can choose a position and time when one of these waves is at a maximum, and define these as $x = 0$ and $t = 0$, respectively. For this wave, we have essentially set its phase constant ϕ equal to zero. There is no guarantee that the other wave will also be at a maximum at this place and time, so we can't arbitrarily set the other phase constant equal to zero as well. If they do have a relative phase that is non-zero, then it turns out that the combined sound will not be as loud as if they waves were in phase, but the general features described below remain the same.

To simplify the math, we will therefore just look at the simple case where the waves are in phase, and set $\phi_1 = \phi_2 = 0$. With the phase constants equal to zero, we now have a simple pair of functions to work with (note that we can drop the minus signs on the times, since $\cos(x) = \cos(-x)$). Superposing these and replacing the periods with frequencies gives:

$$P_{tot}(t) = P_1(t) + P_2(t) = P_o \cos(2\pi f_1 t) + P_o \cos(2\pi f_2 t) \quad (1.3.3)$$

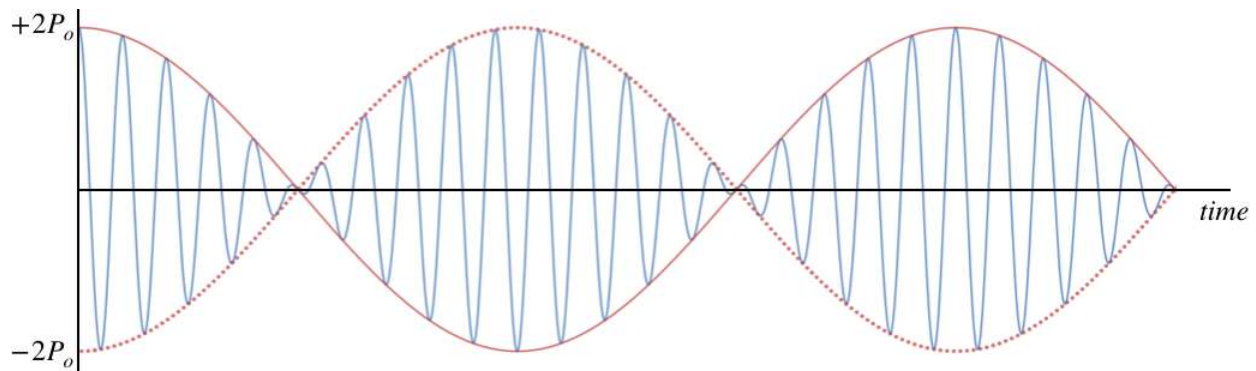
Now apply a trigonometric identity:

$$\begin{aligned} \cos X + \cos Y &= 2 \cos\left(\frac{X+Y}{2}\right) \cos\left(\frac{X-Y}{2}\right) \\ \Rightarrow P_{tot}(t) &= 2P_o \cos\left[2\pi\left(\frac{f_1+f_2}{2}\right)t\right] \cos\left[2\pi\left(\frac{f_1-f_2}{2}\right)t\right] \end{aligned} \quad (1.3.4)$$

This looks like a complicated mess, but there is a reasonable way to interpret it. If we identify the second cosine function as the variation of the pressure that defines the time portion of the sound wave (often referred to as the *carrier wave*), the tone produced has a frequency that is the average of the two individual frequencies. The first cosine function can then be combined with $2P_o$, and together they can be treated as a *time-dependent amplitude*. The amplitude is directly related to the intensity, which is a measure of

the loudness of the sound, so a harmonically-oscillating amplitude would manifest itself as regularly-spaced increases and decreases in volume (beats). It's a bit easier to see how this works with a diagram.

Figure 1.3.5 – Graph of Pressure vs. Time at a Fixed Point Exhibiting Beats



The blue function shown represents the fluctuations in pressure (at position $x = 0$ as a function of time as the sound wave passes through that point. Notice that the peaks have varying heights, reflecting the time-varying amplitude. The red "envelope" function traces just the amplitude of the wave. The intensity is proportional to the square of the amplitude, so the sound is loud everywhere that a red bump occurs, and silence is heard when the red function crosses the axis.

Alert

*The volume of the sound doesn't drop to zero every time the blue graph crosses the axis! That is simply when the pressure happens to be passing through the equilibrium point, and there is still energy in a wave when the displacement of the medium passes through the equilibrium point (in the form of kinetic energy of the medium). But when the **amplitude** of oscillations goes to zero, the energy is zero, and in the case of sound, this means silence.*

We can ask how frequently the beats occur. A beat occurs every time a bump in the red graph occurs. Notice it doesn't matter whether the bump is up or down in the red cosine function, since the intensity is the *square* of the amplitude. Therefore there are two beats are heard for every full period of the red cosine function; one for the bump at the peak, and one for the bump at the trough. This means that the frequency of beats is twice the frequency of the red cosine function:

$$f_{\text{beat}} = 2 \left(\frac{f_1 - f_2}{2} \right) = f_1 - f_2 \quad (1.3.5)$$

[Note: We have assumed that f_1 is the greater of the two frequencies, as negative frequencies have no meaning. One can remove this assumption by defining beat frequency in terms of the absolute value.] We can measure the time spans between peaks for the blue function. This is the period of the carrier wave, which we have already stated is the average of the frequencies of the two sound waves:

$$f_{\text{carrier}} = \frac{f_1 + f_2}{2} \quad (1.3.6)$$

We only really hear the beats clearly when the two individual frequencies are close together – if there is too great of a difference, then the beat frequency is very high, and the beats come too frequently. In this case it's impossible for the human ear to tell when one beat begins and the other ends. In that case, it sounds to the human ear more like a mixture of two sounds – one at the average frequency and one at the much lower "beat frequency."

Example 1.3.3

A standing wave of sound persists within a hollow pipe that is open at both ends. A bug within the pipe walks along its length at a speed of $4.0 \frac{\text{cm}}{\text{s}}$. As it does so it passes through nodes and antinodes of the standing wave, hearing the tone get alternately loud and silent, with the time between moments of silence equaling 2.0s. From what we have learned about standing waves, we can compute the frequency of the tone. The distance between the nodes is one-half wavelength of the traveling sound wave, and based on the ant's speed and the time between nodes, we can determine this distance. Then with the wavelength and the speed of sound, we get the frequency:

$$\text{distance between nodes} = v_{\text{bug}} t = \frac{\lambda}{2} \Rightarrow \lambda = 2 \left(4.0 \frac{\text{cm}}{\text{s}} \right) (2.0\text{s})$$

$$= 16\text{cm} \Rightarrow f = \frac{v}{\lambda} = \frac{344 \frac{\text{m}}{\text{s}}}{0.16\text{m}} = 2150\text{Hz}$$

When asked about the experience, the bug claimed it had no idea that it was walking through loud and silent regions. It described the experience as hearing periodic "beats" every 2.0 seconds in the tone. Re-derive the frequency of the sound from the perspective of the bug.

Solution

This problem is challenging because it's not immediately clear how beats can occur, when beats require two sound waves of different frequencies. There are two sound waves here – one reflected by each end of the pipe – but they both have the same frequency... or do they? They have the same frequency from our perspective, but **not from the bug's!** The bug is moving toward one end of the pipe and away from the other, so one of the waves is doppler-shifted to a higher frequency, and the other wave is doppler-shifted to a lower frequency. These waves of different frequencies interfere to provide beats from the ant's perspective. So now that we see how this makes sense, we need to do some math. Both doppler shifts involve a stationary source (the ends of the pipe) and a moving receiver (the bug):

$$f_1 = \left(\frac{v + v_{\text{ant}}}{v} \right) f \quad f_2 = \left(\frac{v - v_{\text{ant}}}{v} \right) f$$

The frequency f_1 is what is heard by the ant from the wave coming from in front of it, and f_2 is the frequency of the sound coming from behind it, and f is the frequency of the sound from the perspective of the stationary ends of the pipes (i.e. the frequency of the standing wave). The beat frequency heard by the ant is simply the difference of these two frequencies, so:

$$f_{\text{beat}} = f_1 - f_2 = \left(\frac{v + v_{\text{ant}}}{v} \right) f - \left(\frac{v - v_{\text{ant}}}{v} \right) f = 2 \frac{v_{\text{ant}}}{v} f$$

We know that the bug hears silence every 2.0 seconds, so the beat frequency is 0.50Hz . Plugging this, the speed of the ant, and the speed of sound into this equation gives:

$$f = \frac{v}{2v_{\text{ant}}} f_{\text{beat}} = \frac{344 \frac{\text{m}}{\text{s}}}{2 (0.040 \frac{\text{m}}{\text{s}})} (0.5\text{Hz}) = 2150\text{Hz}$$

While we have worked this out for some specific numbers, it is of course true in general. If the bug was moving faster, then the pulses at the antinodes would come more frequently. From the bug's perspective the doppler shifts would greater separate the two individual sound frequencies, leading to a higher difference and faster beat frequency, and this identically equals the rate of passing through antinodes.

CHAPTER OVERVIEW

10: Intrinsic Angular Momentum – "Spin"

[10.1: Measuring Angular Momentum](#)

[10.2: It's Not Rotating!](#)

[10.3: Fermions and Bosons](#)

This page titled [10: Intrinsic Angular Momentum – "Spin"](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

10.1: Measuring Angular Momentum

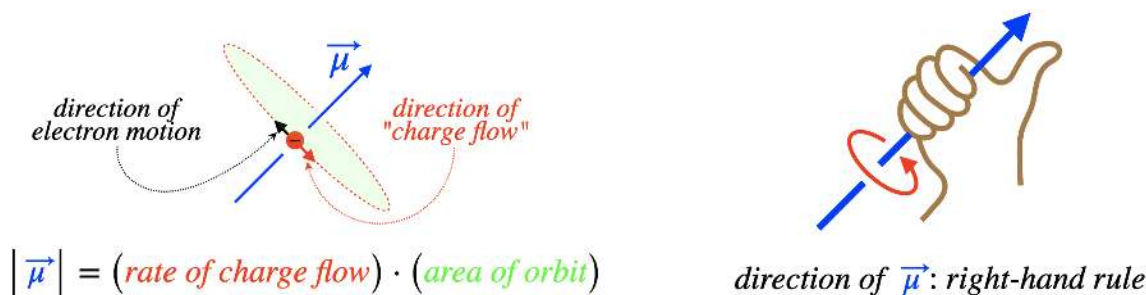
Linking Angular Momentum to Magnetism

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 or 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:

$$|\vec{\mu}| = \left(\frac{e}{\frac{2\pi r}{v}} \right) (\pi r^2) = \frac{evr}{2} \quad (10.1.1)$$

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:

$$\vec{L} = \vec{r} \times \vec{p} \Rightarrow |\vec{L}| = mvr \quad (10.1.2)$$

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:

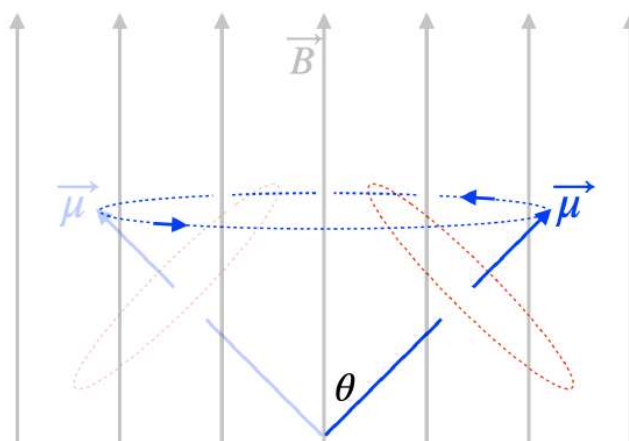
$$\vec{\mu} = -\frac{e}{2m_e} \vec{L} \quad (10.1.3)$$

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:

$$\vec{\tau} = \vec{\mu} \times \vec{B} \Rightarrow |\vec{\tau}| = |\mu| |\vec{B}| \sin \theta \quad (10.1.4)$$

When we studied the gyroscope in Physics 9A, 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:

$$\Omega = \frac{\tau}{L_{\perp z}} = \frac{\tau}{L \sin \theta} \quad (10.1.5)$$

Plugging in for the torque and the angular momentum gives the precession rate, known as the **Larmor frequency**:

$$\Omega = \frac{\mu B \sin \theta}{\frac{2m_e}{e} \mu \sin \theta} = \frac{eB}{2m_e} \quad (10.1.6)$$

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.

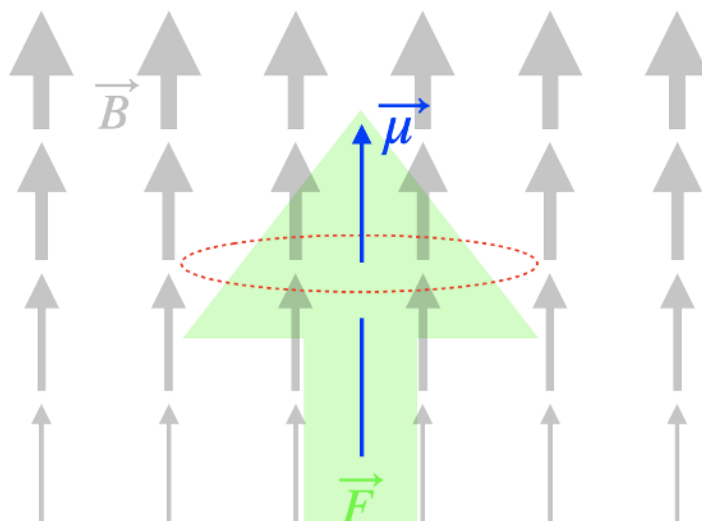
A couple items of note here:

- 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 9.4.1](#). 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:

$$F = \mu \frac{dB}{dz} \quad (10.1.7)$$

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:

$$F = \mu_z \frac{\partial B_z}{\partial z} \quad (10.1.8)$$

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 9.4.12](#)), 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 an 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.

This page titled [10.1: Measuring Angular Momentum](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

- [1.2: Vector Multiplication](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

10.2: 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 through 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 10.1.1](#) – 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 9A, 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*.

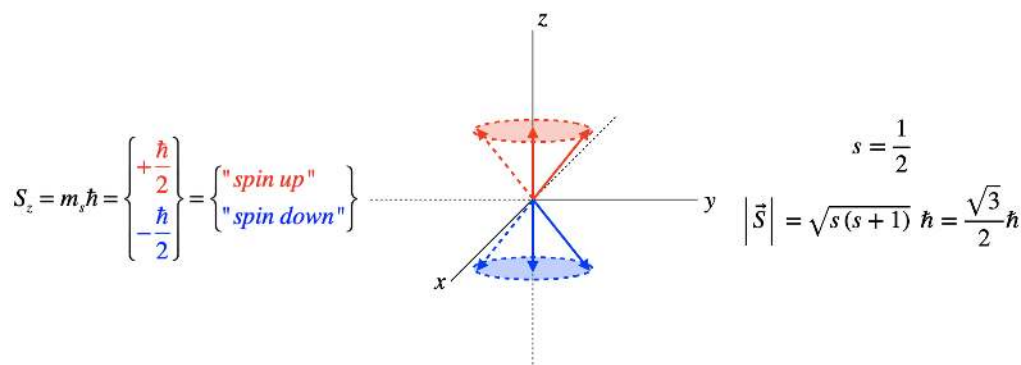
We can do further SG experiments to look into the magnitudes of angular momentum for these two states. For the $l = 1$ orbital case, we found from the deflection of the $m_l = \pm 1$ cases that those z components have a magnitude of $\hbar$. If we compare this to the results

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-1/2*.

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 9.4.2](#):

Figure 8.2.1 – States of Spin-1/2



Gyromagnetic Ratio and g-Factor

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. This relation comes down to a simple ratio of these two quantities, known as the *gyromagnetic ratio*:

$$\gamma = \frac{\text{magnetic moment}}{\text{angular momentum}} \quad (10.2.1)$$

In our previous calculation that related $\vec{\mu}$ to $\vec{L}$, the radii of the circular orbit divided-out, resulting in [Equation 10.1.3](#). 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 *g-factor*:

$$\vec{\mu} = \pm \frac{ge}{2m} \vec{S} \quad (10.2.2)$$

[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:

$$\Psi(x, t) \rightarrow \Psi(x, t) \otimes \chi_s \quad (10.2.3)$$

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 " $\otimes$ " 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:

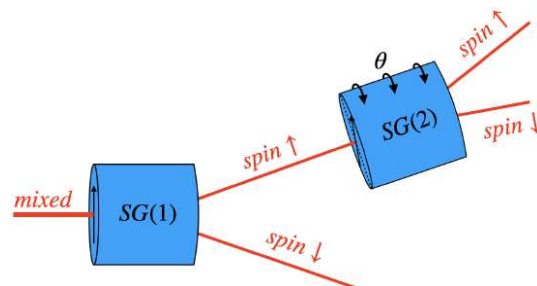
$$\chi_s = \alpha \chi_{\uparrow} + \beta \chi_{\downarrow}, \quad |\alpha|^2 + |\beta|^2 = 1 \quad (10.2.4)$$

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-1/2. 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:

$$\chi_s [\text{after } SG(2)] = \alpha \chi_{\uparrow} + \beta \chi_{\downarrow} = \cos\left(\frac{\theta}{2}\right) \chi_{\uparrow} + \sin\left(\frac{\theta}{2}\right) \chi_{\downarrow} \quad (10.2.5)$$

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- $\frac{1}{2}$ 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:

$$P(\uparrow) = |\alpha|^2 = \cos^2\left(\frac{\theta}{2}\right) \quad (10.2.6)$$

The half-angle argument of the cosine function differs from the full-angle result for Malus's law.

This page titled [10.2: It's Not Rotating!](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

10.3: 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- $1/2$).

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 whole.

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 our 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:

$$\Psi(\text{two-particle system}) = \Psi(\text{particle } A) \otimes \Psi(\text{particle } B) \quad (10.3.1)$$

The real key here is the probabilities. The probability density comes out to be just a product of probability densities:

$$\begin{aligned} P(\text{two-particle system}) &= |\Psi(\text{two-particle system})|^2 \\ &= |\Psi(\text{particle } A)|^2 \otimes |\Psi(\text{particle } B)|^2 \end{aligned} \quad (10.3.2)$$

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 $x_2 + dx_2$:

$$P(A \text{ near } x_1 \text{ and } B \text{ near } x_2) = |\Psi_A(x_1)|^2 dx_1 |\Psi_B(x_2)|^2 dx_2 \quad (10.3.3)$$

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:

$$|\Psi_{\text{system}}(x_1, x_2)|^2 dx_1 dx_2 = |\Psi_A(x_1)|^2 dx_1 |\Psi_B(x_2)|^2 dx_2 \quad (10.3.4)$$

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.

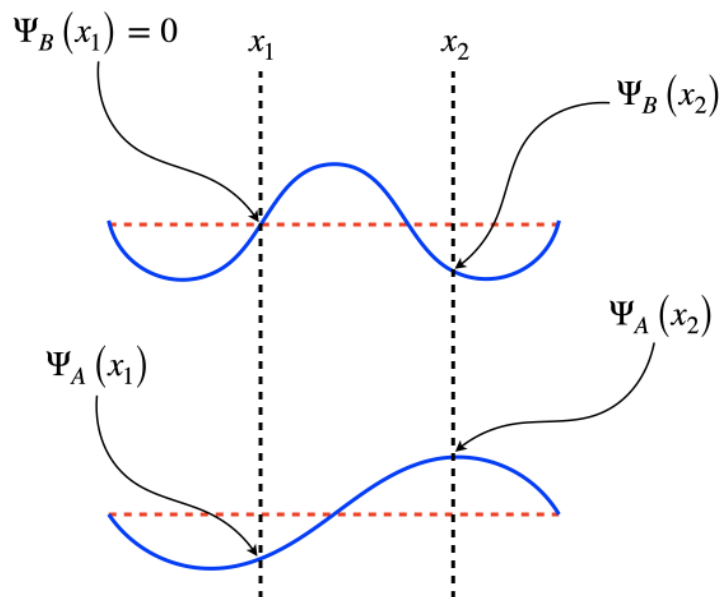
Exchange Symmetry

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:

$$|\Psi_{\text{system}}(x_1, x_2)|^2 dx_1 dx_2 = |\Psi_{\text{system}}(x_2, x_1)|^2 dx_1 dx_2 \quad (10.3.5)$$

It should be clear that this is not the case for Equation 10.3.4, 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 10.3.1 – 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 10.3.4, 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:

$$\Psi_+(x_1, x_2) = \frac{1}{\sqrt{2}} [\Psi_A(x_1) \Psi_B(x_2) + \Psi_A(x_2) \Psi_B(x_1)] \quad (10.3.6)$$

$$\Psi_-(x_1, x_2) = \frac{1}{\sqrt{2}} [\Psi_A(x_1) \Psi_B(x_2) - \Psi_A(x_2) \Psi_B(x_1)] \quad (10.3.7)$$

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.

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.

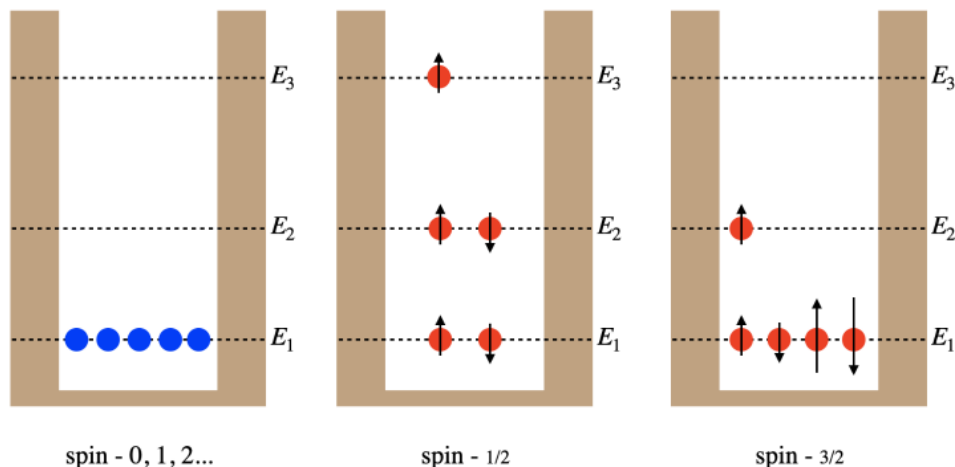
Let's take as an example a one-dimensional infinite square well (we aren't very creative, this is the first example we use for everything). Let's place five identical particles into this "box," such that they reach an energy eigenstate. [Keep in mind we are talking about a multi-particle quantum state 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 10.3.2 – Multi-Particle Ground State According to Particle Spin



This page titled [10.3: Fermions and Bosons](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

CHAPTER OVERVIEW

2: Foundations of Special Relativity

[2.1: The Relativity Principle](#)

[2.2: The Nature of Time](#)

[2.3: More Thought Experiments](#)

[2.4: Paradoxes](#)

This page titled [2: Foundations of Special Relativity](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

2.1: The Relativity Principle

Inertial Frames

Back in our studies of classical mechanics, we spent a very brief period of time learning about how to relate the measurements of position and time between two observers in relative motion (go [here](#) for the LibreText reminder of this topic). Actually "relative motion" in this context is imprecise. We restricted our study to a very specific kind of relative motion – that for which the two observers maintain a constant relative velocity.

For what is to come, we will restrict these frames of reference even further – we will insist that every observer makes its measurements from an *inertial frame*. This kind of frame is one in which Newton's first law assures that objects will not spontaneously begin to accelerate. That is, if we are in such a frame, and we eliminate all of the real forces present on a stationary object, then the object remains at rest in that frame.

The simplest way to understand inertial frames is to consider what kind of frame is *not* inertial. Suppose you are in spaceship (far away from all gravitational sources), and it is accelerating forward. You hold a pencil in your hand, which is at rest in your frame, but you are exerting a force on it with your fingers, so to test to see if you are in an inertial frame, you release it. As soon as you do, it continues with whatever speed it had at the moment of release, while you and your spaceship continue to accelerate. From your perspective, it is the pencil that accelerates, which tells you that you are not in an inertial frame.

[One might wonder why an inertial frame is an additional restriction beyond what we did in 9A. Certainly the intention was to deal with inertial frames back then, but technically two frames with equal acceleration vectors (but unequal velocities) will also satisfy the Galilean transformation.]

Postulate(s)

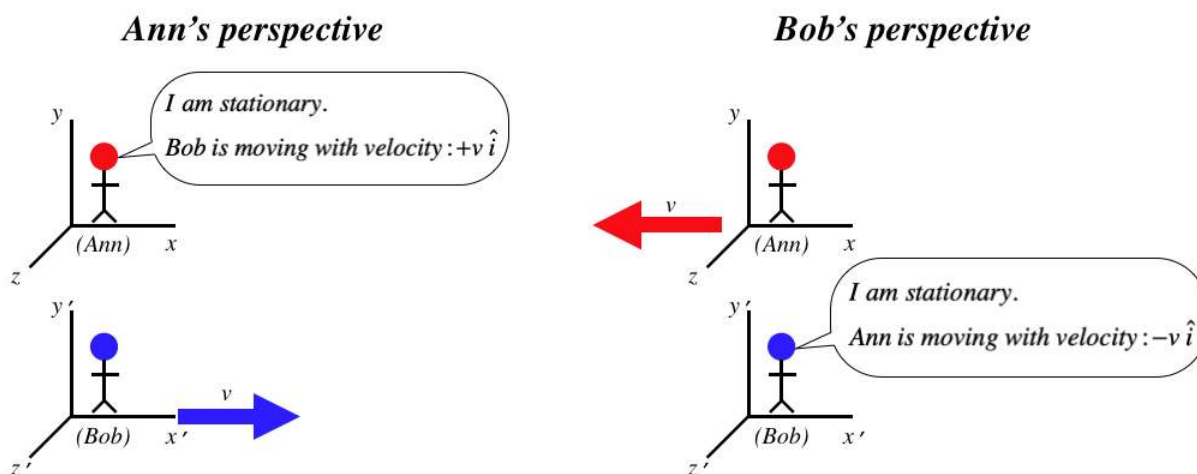
We have a simple experiment for testing whether our frame is inertial, but it doesn't tell us whether our frame is stationary or moving in a straight line at a constant speed, because when we release the pencil under these circumstances, it remains stationary from our perspective in both cases. So what kind of experiment will tell us whether or not we are moving?

Albert Einstein pondered this very thought, and came up with no answer. Eventually, he felt compelled to assert it as a fundamental aspect of our universe, and the *relativity principle* was born:

No experiment can be performed within an inertial frame that determines whether it is moving or at rest.

This is also known as the first postulate of the theory of Special Relativity. One way that we can express this is in terms of an "argument" between two observers.

Figure 1.1.1 – All Observers in Inertial Frames Can Claim to Be Stationary.



[These kinds of diagrams, where the perspectives of two observers in relative motion, will have some common elements. First, we will always define their relative motion to be parallel to their common x -axes. Second, we will define the primed frame to be moving in the $+x$ direction relative to the unprimed frame.]

Calling this the "first postulate" implies that there is a second postulate, and there is, though one could argue that it follows directly from the first postulate and therefore doesn't need to be stated separately. It is this:

Every observer measures the velocity light to be the same value.

The reason this "second postulate" can be considered a consequence of the first postulate is that the theory from which we derive the speed of light contains no provisions for the motion of the observer (or rather, it predicts the same speed for all observers). Therefore the theory predicts that any experiment that measures the speed of light in a vacuum will give a specific answer. If different inertial frames produced different values for the speed of light, we would have a violation of the first postulate, as we would then have an experiment to determine the "true" rest frame. So to the extent that we accept this theory of light propagation, we don't need the second postulate.

Digression: "Ultimate Speed"

The discussion above actually paints a somewhat inaccurate picture of the foundation of relativity. As we will see later, these postulates lead to the requirement of the speed of light being the limit which no relative motion can ever exceed. It turns out that if we just postulate that such an "ultimate speed" exists, then relativity results, independent of the theory of light propagation. That is, light sort of "coincidentally" travels at the ultimate speed, but the theory of relativity would apply even if it didn't, so long as this "cosmic speed limit" exists.

A Bit About Waves

The first postulate seems innocuous enough, and perhaps it even seems intuitive. But this business about every observer measuring the same speed for a beam of light didn't sit well with many physicists at the time Einstein proposed it. To see why, and to fully understand the implications of this being true, we need to review a little bit about waves, because as we know from Physics 9B and 9C, light is a wave phenomenon. Of all the things we previously learned about waves, these are the properties of waves we will most need to recall for this discussion...

effect of medium

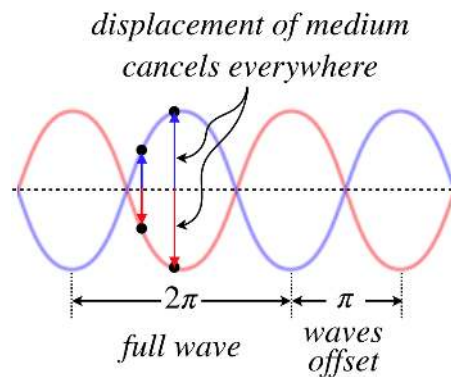
Waves are phenomena which transfer energy through space by the means of a self-propagating disturbance of a medium (the stuff that is "waving"). A wave on a string carries energy along because a piece of the string displaces, which pulls on an adjacent piece of string, displacing it, and so on. Surface waves on water and sound waves through air work in a similar manner. What these all have in common is that medium itself doesn't travel with the energy (it moves with the disturbance), and the speed with which adjacent particles in the medium interact with each other is a property of the medium. Put another way, the speed of a wave is exclusively a function of the properties of the medium. A wave on a string travels faster when the string is more taut, and slower when the string is made is more dense, for example.

superposition

When two waves traveling through the same medium encounter each other, the effects they have on the medium are additive at that the same point in the medium. So if one wave displaces a string by y_1 at some position x on the string, and another wave displaces the string by y_2 at the same position, then the total displacement of the string with the two waves present at x at the same time is $y_1 + y_2$. This additivity property is called *superposition*, and it has the particularly interesting feature that two waves can actually *cancel* each other entirely (a phenomenon called *destructive interference*), if the two waves happen to be displacing the medium equal amounts in opposite directions.

If a wave happens to be harmonic (the medium displacement as a function of position and time is a sine wave), then this destructive interference occurs when two identical waves are out of phase by π .

Figure 1.1.2 – Destructive Interference of Two Harmonic Waves



The Michelson-Morley Experiment

The crux of the problem for those originally opposed to Einstein's assertion that the speed of light is the same when measured by any observer is that *this is not true of other waves*. If one moves through the air into an oncoming sound wave, that sound wave is moving faster relative to that person than relative to someone stationary in the air. The point is that, as stated above, the speed of a wave is entirely determined by the medium, and this speed is *relative to that medium*. So if an observer moves relative to a medium, then the relative speed of waves propagating through that medium can change.

So we are left with the question: What medium is disturbed as a light wave passes through it, and can't we see a change in the speed of light if we just move relative to this medium? This was an open question at the time of Einstein, and was very puzzling for a couple reasons. The first is that the theory of light did not require the existence of any medium at all.

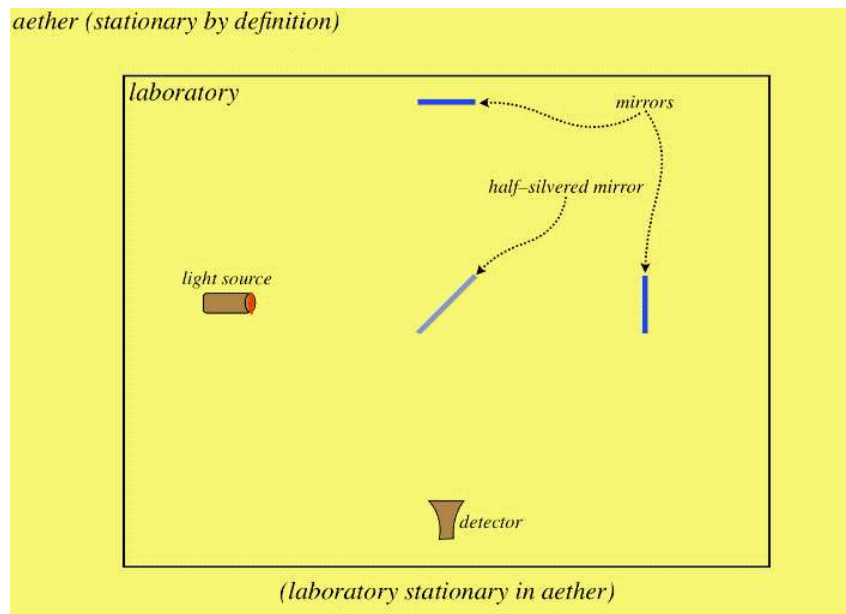
But far more puzzling than this was the fact that no one could seem to come up with any experimental evidence of the existence of a medium for light (which they referred to as the *luminiferous aether*). The most convincing null result of such a search was performed by two American physicists Albert Michelson and Edward Morley, who employed a device called an *interferometer*, the basics of which are still used today for countless applications.

The basics of an interferometer work like this: A single light beam is split into two separate beams. As they came from the same source, they are in phase with each other, but if we send them on separate journeys, and then bring them back together, they may no longer be in phase. For example, one of the beams may travel farther than the other. If the difference in their phases when they get back together is just right, they will destructively interfere with each other.

[For the two beams to behave this way, the original beam needs to be somewhat coherent, meaning that parts of the light near each other are in phase. Nowadays we easily achieve this artificially with lasers (sunlight is also quite coherent), but fortunately the degree of coherence achievable in a lab in the late 1800's was sufficient to successfully perform this experiment.]

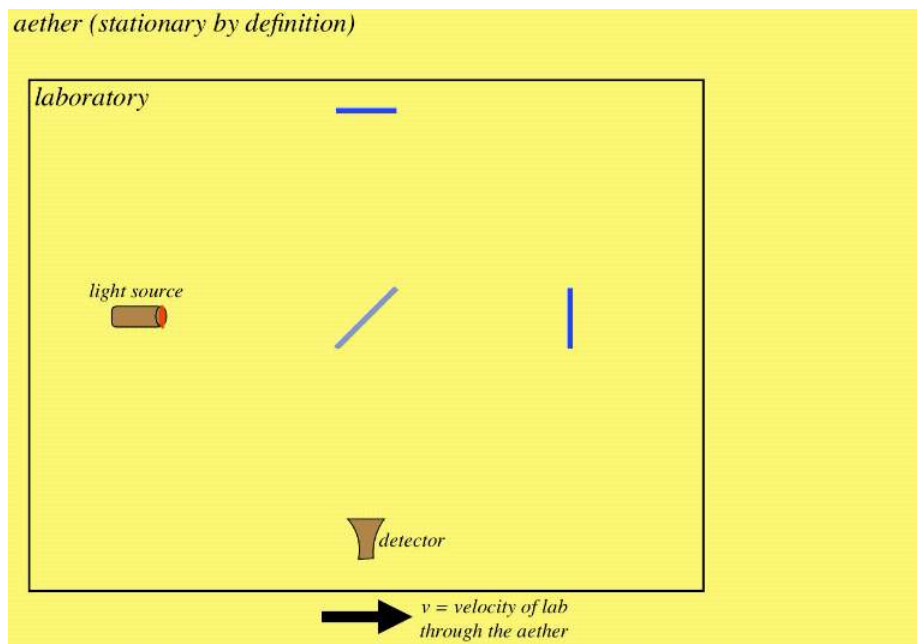
The ingenuity of the MM interferometer is that it splits the beam of light and sends the two pieces on journeys that are equal distances, but are at right angles to each other. If the laboratory is stationary in the aether, then the beams come back to the detector in phase, and do not cancel each other out:

Figure 1.1.3 – Laboratory Stationary in Aether



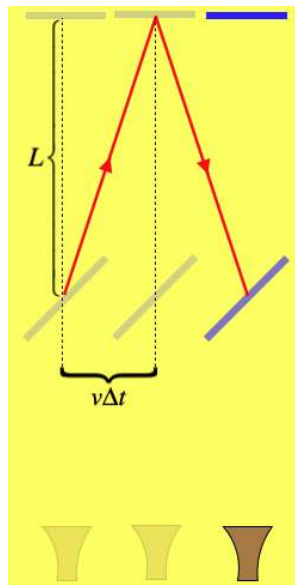
If this device is moving through aether (let's say in a direction parallel to one of these directions), then one of the arms of the interferometer moves parallel to the direction of the light during the journey, while the other moves perpendicular to it. These two journeys are not equal in distance, which means the beams can enter the detector out of phase:

Figure 1.1.4 – Laboratory Moving through Aether



We can do a bit of math to determine the difference in the distance between these two journeys. The only difference between the paths of the two beams comes after they encounter the half-silvered mirror the first time, and before they encounter it the second time, so we focus on this portion of the process for both of them:

Figure 1.1.5 – Distance Traveled by the Transverse Beam



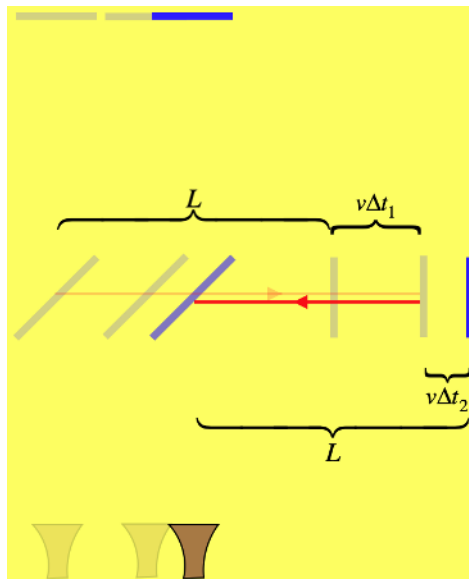
The transverse beam (the one perpendicular to the laboratory's motion) follows a diagonal path in both directions, the distance of which can be determined from the Pythagorean Theorem:

$$s = 2\sqrt{L^2 + (v\Delta t)^2} \quad (2.1.1)$$

But the distance traveled by the beam also equals the speed of light through the aether c multiplied by the time elapsed, so we get:

$$s = 2c\Delta t \Rightarrow s = 2\sqrt{L^2 + \left(v\frac{s}{2c}\right)^2} \Rightarrow s = \frac{2L}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (2.1.2)$$

Figure 1.1.6 – Distance Traveled by the Longitudinal Beam



The longitudinal beam (the one parallel to the laboratory's motion) stays on a straight line, but the time it spends going in each direction is not the same, so we compute these times separately then use them along with the speed of light to get the full distance this beam traverses during this part of the process:

$$\left. \begin{aligned} s_1 &= L + v\Delta t_1 = c\Delta t_1 \Rightarrow \Delta t_1 = \frac{L}{c - v} \\ s_2 &= L - v\Delta t_2 = c\Delta t_2 \Rightarrow \Delta t_2 = \frac{L}{c + v} \end{aligned} \right\} \Delta t_1 + \Delta t_2 = \frac{2LC}{c^2 - v^2} \quad (2.1.3)$$

$$\Rightarrow s = c(\Delta t_1 + \Delta t_2) = \frac{2L}{1 - \frac{v^2}{c^2}}$$

We can see from these two results that the two beams travel different distances. The denominator for each case is a fraction that is less than 1, so the square root is the larger denominator. This means that the longitudinal path (with the smaller denominator) is longer than the transverse path. The difference between these paths can be expressed by a multiplicative factor that depends upon the speed of the laboratory through the aether and the speed of light:

$$s_{longitudinal} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} s_{transverse} \quad (2.1.4)$$

We therefore find that if the laboratory is moving through the aether at just the right speed, then the difference in distances traveled will equal half the length of a wave, and the two beams will cancel at the detector, resulting in darkness. In fact, this is not the only case that will give this result. If the difference in distance traveled is $\frac{3}{2}$ of a full wave, then again the troughs and crests of the waves will match up, making destructive interference. Moreover, even if there is not complete destructive interference, every offset of the waves will result in some variation in the brightness of the light.

So given that the earth is moving through space, and is also rotating, one would expect that it is not difficult to find some evidence that the aether exists, and that we are moving through it. But try as they might, Michelson and Morley found no such thing. Many explanations were offered for the problem (many before the experiment was even performed), such as the aether being "dragged" by the earth, so that it was stationary around us here on the surface (logical arguments based on starlight observations and additional experiments on starlight proved this to be false). Einstein alone showed the courage to just discard the existence of the aether altogether, no matter how nonsensical the consequences might seem to be at first.

This page titled [2.1: The Relativity Principle](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

2.2: The Nature of Time

Spacetime Events

We now embark on deriving the consequences of the relativity principle in the same way that Einstein did – using a tool he called *Gedankenexperiment* (*thought experiment*). In order to keep everything straight in our discussions, we begin by defining a *spacetime event*.

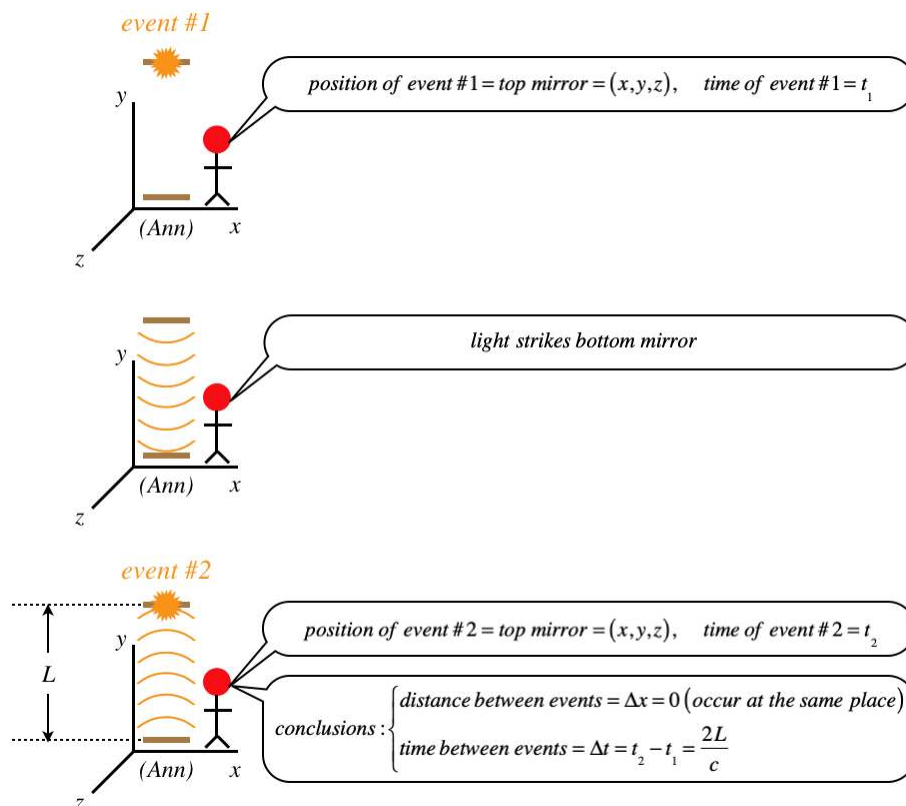
In the context of special relativity, a spacetime event is an instantaneous occurrence at a specific point in space and at a specific moment in time. A single point on a stationary light bulb as it dims defines a specified location, but it is not an event because the dimming process does not occur at a single instant in time. A baseball bat at exactly 12:01pm occurs at a single instant in time, but it is not an event, because the position is not specified at a single point. An easy way to visualize a spacetime event is to picture it as a very quick flash of light from a point source. The position of the point source and the instant in time the flash occurs define the space and time “coordinates” of the spacetime event.

It is much easier to define what a spacetime event is than it is to put physical quantities in terms of the spacetime coordinates, but as we will see, this is exactly what we will have to do to make sense of what is to come. We begin with one of most startling results, which is ironically one of the easiest to derive.

Time Dilation

Our first thought experiment involves turning the function of a clock into a series of spacetime events. This clock functions as follows: Light bounces back-and-forth between two mirrors, and every time it strikes one of the mirrors, the clock “ticks.” We begin with Ann's perspective on what is happening with this clock. She happens to be in the same frame as the two mirrors, so to her they are at rest, and the light is bouncing parallel to her y -axis. The two spacetime events we will look at are two consecutive ticks of the clock.

Figure 1.2.1 – Ann's Perspective of the Light Clock

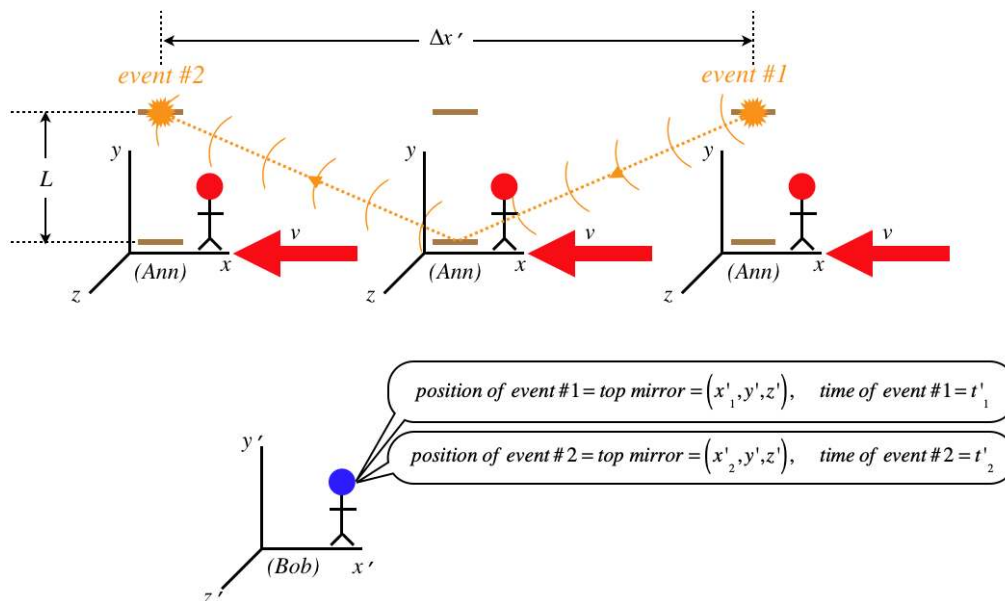


Okay, so we have used the two events to determine the time span between them according to Ann. The goal of relativity is to describe what a second observer measures for a physical process given what the first observer measured. So now we introduce Bob,

who is in what we call the primed inertial frame, moving at a constant speed v in the $+x$ -direction relative to Ann. One might interject, "Wait, this is *time* we are talking about! Won't both of them measure the same amount of time between ticks of the clock?" Don't assume anything in relativity – just use the spacetime events and the postulate(s), and see where it leads.

Looking from Bob's perspective means that not only is Ann moving in the $-x$ -direction, as we noted previously, but the two events (which both occur at the top mirror) don't occur at the same position in space, since the mirror moves:

Figure 1.2.2 – Bob's Perspective of the Light Clock



Now we calculate the time between the two events, as we did for Ann. From Bob's perspective, the light travels a longer distance than Ann measures, and very importantly, *both Ann and Bob measure the speed of light to be the same* (postulate of relativity), so Bob must measure a longer time period than Ann measures between the same ticks of the light clock! According to Bob, the light travels diagonally from the top mirror to the bottom one, and the length of this half of the trip can be written in terms of the speed of light, and in terms of the Pythagorean Theorem:

$$\begin{aligned}\Delta x' &= x'_2 - x'_1 = v\Delta t' \\ c\Delta t' &= 2\sqrt{L^2 + \left(\frac{\Delta x'}{2}\right)^2}\end{aligned}\quad (2.2.1)$$

We can eliminate $\Delta x'$ from these two equations to relate the time span measured by Bob to the time span measured by Ann:

$$\Delta t' = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \left(\frac{2L}{c} \right) = \gamma_v \Delta t, \quad \gamma_v \equiv \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}\quad (2.2.2)$$

The time between ticks for Bob is greater than the time between ticks for Ann by a factor of γ_v (which is clearly a constant greater than 1). Just to clarify, this is not an optical illusion for Bob – he doesn't just "see the clock ticking slower than it really is," *it is actually ticking slower*. Also, it is important to note that while we used light to achieve this answer, it doesn't just apply to light phenomena, it applies to time flow in all its manifestations. If Ann measures her own pulse to be 60 beats per second (one second between each beat), and $\gamma_v = 2$, the Bob would measure Ann's heart rate to be 30 beats per second (2 seconds between each beat).

It's worth taking a moment to review what the source of this result is. It comes from the fact that the light in the light clock travels farther for Bob than it does for Ann, but they agree on the speed of that light, which means that the time between the two events must be greater for Bob than it is for Ann.

As startling as this result is, it gets weirder. Suppose Bob has a light clock exactly like Ann's. What does Ann observe when she looks at Bob's clock? She sees exactly the same thing happening with Bob's clock as he sees with her clock! Therefore Ann claims that time is passing slower for Bob than it is for her, even as Bob says that Ann's time is passing slower than his own. Which one of them is correct? Is Ann's time passing slower, or is Bob's? They are both in inertial frames, so according to the principle of relativity, each has an equal right to declare themselves to be "stationary." Therefore they are both right. The reason it seems like it

is impossible that this can be true is that we cling to the incorrect notion that time is universal. The time span between two events is a relative quantity that depends upon who measures it.

Recording Spacetime Coordinates

While the calculation above is correct, it does require an assumption that we need to briefly address. Both Ann and Bob noted the positions and times of the events in their frames. Given the importance of both position and time in relativity, we need to be specific about how these numbers are recorded. What is observed is a spacetime event, which we have modeled as a flash of light that occurs in an instant at a specific position. So let's imagine constructing a massive lattice of labeled positions throughout all of space, and the position of any possible flash must coincide with one of those positions, giving us our spatial label. Note that every inertial observer can create such a lattice independent of every other observer, because according to the relativity principle, everyone has an equally valid claim to being "stationary." It is true that Bob's lattice of position labels is moving according to Ann, but Ann and Bob only use their own stationary labels to describe the positions of events they see.

To get a complete reckoning of an event, we need to record not only its position, but the time at which it occurs. Given what we know about the rate of time flow for moving clocks, we have to be very careful about how we measure the time at which an event occurs. The one way to be safe is to have the clock that reads the time be positioned at the same place in space as the event. So whenever an event occurs, one simply reads the label of the lattice point at which it occurs, and the value indicated by a clock located at that lattice point when the event occurs.

Two Different Time Measurements

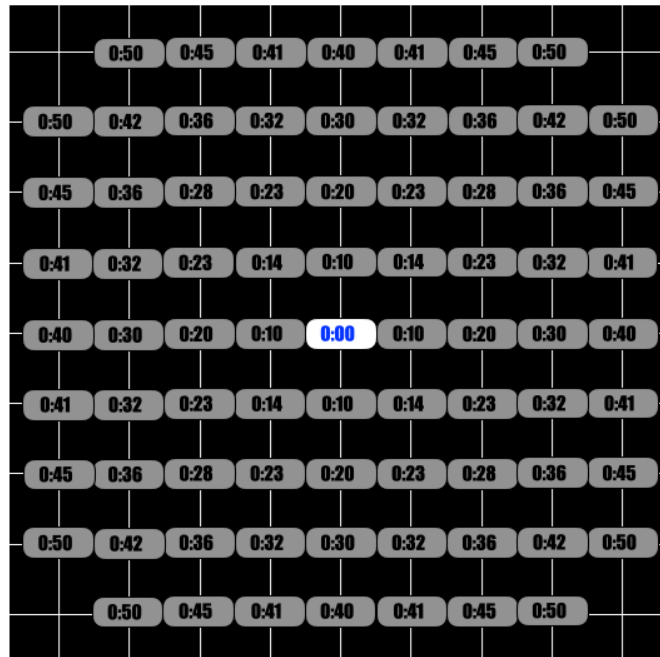
Now that we have a plan for recording data for events in spacetime, we need to give a little more thought to how we plan to have a clock that is properly positioned to measure the time. It turns out that there are two fundamentally different ways to achieve this.

coordinate time

The first way that comes to mind for measuring the time of any given event is to simply place a separate clock at every lattice point. While this is a simple way to get a measurement for any event, we will be interested the time intervals *between two events*, which means that all of our clocks positioned throughout space need to be synchronized. How do we do this? If we bring all our clocks together in one place, set them the same, and then move them out to their assigned locations, then the weird effects that come from relative motion of clocks may un-synchronize them when they are moved. Instead what we can do is this:

1. Distribute all of the clocks throughout space.
2. Set the clock at the origin to a time of 0:00.
3. Using the lattice positions of the clocks, compute the time it will take a spherical wave pulse of light that starts at the origin to reach all the other clocks, add this time to 0:00, and set the clock at this time.
4. Start the clock at the origin while starting the spherical pulse of light from the origin.
5. When the light wave reaches a clock, start it running.

Figure 1.2.3 – Synchronizing Clocks Distributed in Space



By anticipating what the time on the origin clock must be when the light arrives, we can assure that the spherical wave propagates clock synchronization throughout space. This measurement of the time of an event is called *coordinate time* t .

In the example above, both Ann and Bob measured the time between ticks in their own coordinate time. For Ann, the coordinate time span between the two spacetime events was $\Delta t = t_2 - t_1$, while for Bob it was $\Delta t' = t'_2 - t'_1$. As we found in the thought experiment, these values are not equal, which is to say that this manner of measuring a time span between two events is *relative*. Whenever the value of a physical quantity is different when measured from different inertial frames, we say that such a quantity is *frame-dependent*. We therefore declare:

Coordinate time spans are frame-dependent.

proper time

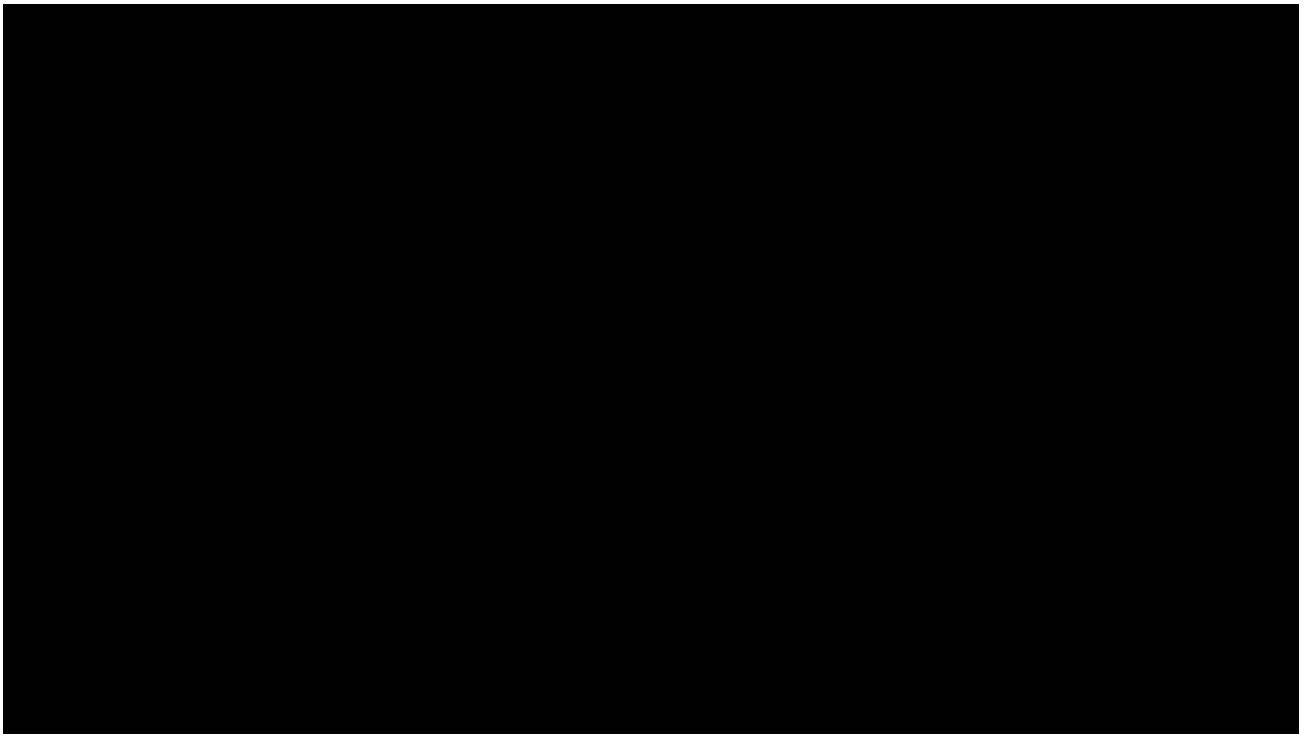
We certainly are not *required* to measure time between events by placing synchronized clocks at all the lattice points in our frame. Another way would be to use a *single* clock that is moved from the lattice point of the first event to the lattice point of the second. As before, a clock records the time of the event while it is at the same point in the lattice as the event, but this time it is the *same* clock, which means we do not need to rely upon our synchronization method above. A time interval measured in this manner is called a *proper time* $\Delta\tau$ between the spacetime events.

Alert

The name "proper time" dates back to the early days of relativity, and is still used today, but it is dangerously misleading for those new to the subject. The word "proper" can easily be misconstrued to mean "correct," and hopefully this section is making it clear that this is cannot be the case. We are in the process of defining two different ways of measuring the time between two events (which can give different answers), and neither of these is any more correct than the other. The sooner the reader purges from their thoughts the notion that time is absolute and that there must be one correct value for the time between two events, the better.

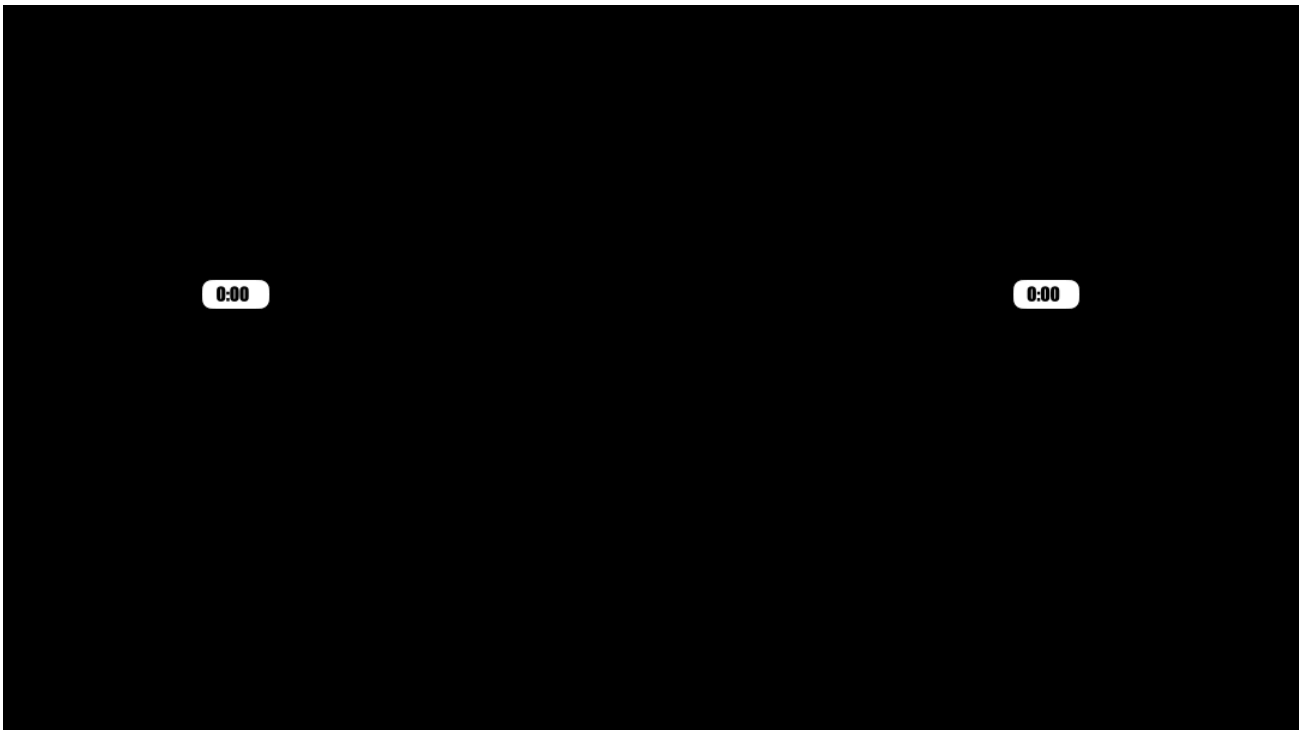
It might seem like both the coordinate and proper time methods of measuring time intervals between events should produce the same result, but in fact they do not. The figures below demonstrates these two measurements for the same two spacetime events.

Figure 1.2.4 – Spacetime Events



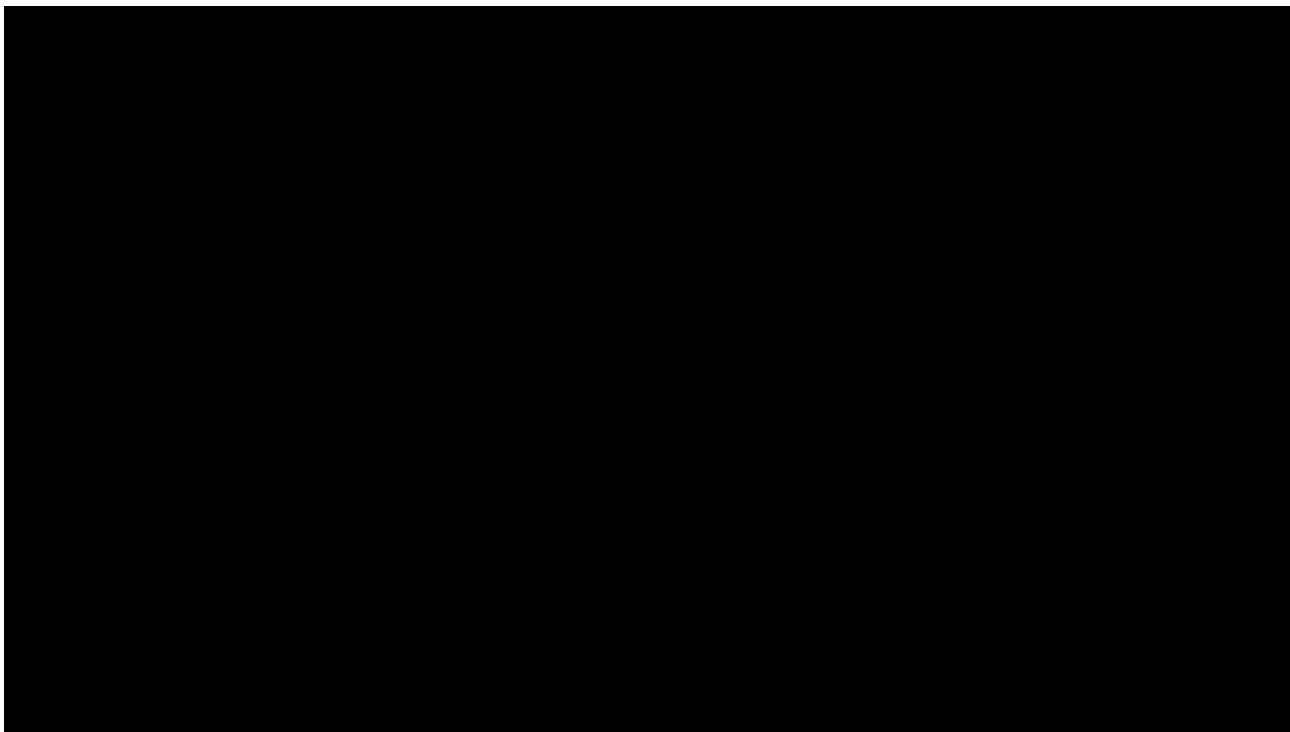
Here we have just two spacetime events viewed from a particular reference frame.

Figure 1.2.5 – Coordinate Time Interval



These are the same two events, viewed from the same reference frame, but the coordinate time clocks placed at the positions of the events are in place to measure the times at which the events occur.

Figure 1.2.6 – Proper Time Interval



Again we have the same two spacetime events, viewed in the same frame, but a clock at rest in a different frame is now visible, and it measures the time interval between the two spacetime events, *but in that frame the events occur at the same position* (at the nose of the rocket ship).

Our example with Ann and Bob earlier shows why these time measurements come out different, if the postulate of special relativity about the constancy of the speed of light is accurate. The two flashes occur at the same lattice point in Ann's frame (the top mirror remains at the same place in Ann's labeled lattice), so she measures the time interval using the proper time method. Meanwhile, the top mirror moves from one lattice point in Bob's frame to another, so he relies upon the synchronized clocks positioned at those points for the time interval. The thought experiment demonstrates through the postulates of relativity that these two time intervals are not equal.

Example 2.2.1

In the figures above that depict two spacetime events and a spaceship moving between them, we are observing from an inertial frame. How fast is the spaceship moving relative to this frame?

Solution

The two events occur at the same position in space in the ship's frame (at its nose - remember that the people on the ship can claim that the ship is not moving, so the two flashes occur at the same place). Therefore the ship measures the proper time interval between the two flashes, just as Ann measured the proper time interval of the two flashes at the top mirror. Using the formula we derived to express the relationship between the time intervals, we find:

$$\Delta t = \gamma_v \Delta \tau \Rightarrow \sqrt{1 - \frac{v^2}{c^2}} = \frac{\Delta \tau}{\Delta t} = \frac{12s}{15s} = 0.8 \Rightarrow v = 0.6c$$

The feature that best distinguishes proper time from coordinate time is the fact that a coordinate system is not needed to measure proper time. For example, we could introduce several other inertial frames of reference to look at the time interval between those top mirror flashes, but *only* Ann's clock will measure the time interval as the one where the flashes occur at the same position in space – every frame other than Ann's will be similar to Bob's, in that the flashes occur at different lattice positions for their frame. All of the observers will agree on one thing – that Ann's measurement of the time interval is somehow "special", and this gives them a way to all agree upon a time interval. Put another way, the measurement of proper time (essentially asking Ann what answer she got, as she was the only observer in an inertial frame to have both events occur at the same position) gives the same result for all reference frames. That is:

Proper time is frame-independent.

Besides "frame-independent," a word typically used to describe a physical quantity like proper time that doesn't vary from one frame to another, is *invariant*.

Note that it is possible for a proper time measurement to be equal to a coordinate time measurement. For example, in the case discussed above, Ann sees the two events occur at the same lattice point in her frame, so if she looks at the clock placed there, it is the same clock measuring the time for both events, which means it also records the proper time. Bob's measurement of coordinate time, on the other hand, is not the proper time, since he reads the numbers off two different clocks – one placed at the lattice point of the first spacetime event, and one placed at the second. From the light clock example, it should be clear that the shortest distance the light has to travel between the two mirrors occurs in Ann's frame. That is, *every* frame other than the "proper frame" that measures coordinate time is going to measure a longer time interval between the events than the proper time.

two important notes

There are two details that have not been tied-up above that we will mention here and address in a future section:

1. The proper time between two events may not be definable, if the events are separated by a distance that is too great for the spaceship to cross in the interval between their occurrences. For example, if the two events viewed in the frame of the figure above occurred simultaneously in that frame, then there is no way for the spaceship to traverse the distance between the events fast enough to allow both events to both occur at the nose of the ship. We will see below that there is another invariant quantity that applies to any two events, and that the proper time interval is a special case of this invariant when an inertial frame exists that can measure the events at the same position in space.
2. We have defined the proper time here as being measured in an inertial frame, but the spaceship could also have both events occur at its nose when it accelerates between the two events. This will come out to a different result than the inertial frame case, so it is important when declaring (as we did above) that the "proper time is frame-independent", that we keep in mind that this is restricted to the specific history of the clock that measures it. That is, every observer will agree on the proper time measured by a single clock that is present at the positions of both events. "Invariance" pertains to different observers, *not* different clocks. A second clock that is present at the positions of both events will not necessarily measure the same proper time interval as the first clock. The way that each clock gets from the first event to the second (namely, its acceleration during the trip) determines how these two proper time measurements may differ.

Cosmic Speed Limit and the Spacetime Interval

When we look back at the time dilation result we obtained above, an obvious question comes to mind: If observe a clock in a moving frame to tick more slowly than one in our rest frame by a factor of $\sqrt{1 - \frac{v^2}{c^2}}$, then what happens when the relative speed of the two inertial frames reaches or exceeds $v = c$? Clearly the result gives a nonsense answer, and while this is far from "proof," we will take this moment to make a declaration that we will later see to be true in many other cases...

Two inertial frames can never have a relative speed that exceeds the speed of light, and this cosmic speed limit can only be attained for light itself.

Technically, there are other phenomena besides light that can propagate at light's eponymous speed, and the criterion for this is a simple one, but we will save that discussion for later. For now, we will generalize the "cosmic speed limit" to state that no "influence" or "information" can be passed from one point in space to another at a speed faster than light can traverse the same distance.

This speed limit gives us a new perspective on time intervals between two events. We said above that there was no way to measure the proper time interval between events that are separated in space and are simultaneous, because there is no way for a single clock to get from one event to the other in time. Now we see that the two events don't need to be simultaneous for this to be true. Because of the cosmic speed limit, there is no way for the proper time interval between two events to be measured if they are separated by enough distance that light cannot travel from the earlier spacetime event to the later one. If light can't make it in time, then neither can a clock, and the proper time cannot be measured.

Let's say that the events occur at positions (x_1, y_1, z_1) and (x_2, y_2, z_2) , and times t_1 and t_2 (with $t_2 > t_1$), as measured in some arbitrary inertial frame, respectively. Then there will be a well-defined proper time measurable by a moving clock if the distance between them is less than the distance that light can travel in the time interval $\Delta t = t_2 - t_1$:

$$\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} < c\Delta t \quad (2.2.3)$$

We can therefore invent a sort of "discriminant" that tells us whether two events can be connected in this way:

$$\Delta s^2 \equiv c^2 \Delta t^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2) \quad (2.2.4)$$

When $\Delta s^2 > 0$, it is possible to move a clock from the earlier event to the later one, so that the clock measures the proper time, otherwise one cannot do this. This quantity Δs^2 is called the **spacetime interval** between the two events.

We said that the quantities Δt , Δx , Δy , and Δz are measured in any arbitrary inertial reference frame, but for a moment let's suppose that the events are sufficiently close together, and look at the value of the spacetime interval in the frame where the events are at the same position (i.e. the frame with the clock that, when viewed by someone else, is moved from the earlier event to the later one). In this frame, $\Delta x = \Delta y = \Delta z = 0$ and Δt is the proper time, which means:

$$\Delta s = c \Delta \tau \quad (2.2.5)$$

So the spacetime interval between two events is just proportional to the square of the proper time interval between those events. We stated earlier that the value of $\Delta \tau$ is an invariant – it is the same when measured in any reference frame. This means that all observers will agree on the spacetime interval between two events that allow for a proper time. But now that we are talking about an abstract mathematical quantity instead of a span of time, we can make the more general statement for *all* pairs of events:

The spacetime interval between any two events is an invariant.

Yes, sometimes this interval is positive (allowing for a measurable proper time interval), sometimes it is negative (not allowing this – the proper time interval is imaginary, whatever that means), and sometimes it is zero (making the proper time interval zero). But whatever it comes out to, the value of Δs^2 is measured to be the same quantity in all frames of reference. Let's be clear about what this means: The values of Δt , Δx , Δy , and Δz are all different for various reference frames, but the special combination of these quantities that equals Δs^2 comes out to be the same in every frame, provided the same two events are involved. In other words, if we have two different reference frames, unprimed and primed, then they measure the same spacetime interval:

$$\Delta s^2 = \Delta s'^2 \Rightarrow c^2 \Delta t^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2) = c^2 \Delta t'^2 - (\Delta x'^2 + \Delta y'^2 + \Delta z'^2) \quad (2.2.6)$$

Example 2.2.2

Show that the spacetime interval of the light flash events at the top mirror in the time dilation thought experiment above are the same for Ann and Bob.

Solution

The flashes occur at the same place in space for Ann, so the spacetime interval in her frame is easy to calculate.

$$\Delta s_{Ann}^2 = c^2 \Delta t_{Ann}^2 - (\Delta x_{Ann}^2 + \cancel{\Delta y_{Ann}^2} + \cancel{\Delta z_{Ann}^2})$$

The flashes occur at different x -values for Bob (but the same y and z positions), so:

$$\Delta s_{Bob}^2 = c^2 \Delta t_{Bob}^2 - (\Delta x_{Bob}^2 + \cancel{\Delta y_{Bob}^2} + \cancel{\Delta z_{Bob}^2})$$

The separation of the two events for Bob is just the speed at which Ann is moving, multiplied by the time interval that Bob measures, so:

$$\Delta x_{Bob} = v \Delta t_{Bob} \Rightarrow \Delta s_{Bob}^2 = c^2 \Delta t_{Bob}^2 - (v \Delta t_{Bob})^2 = (c^2 - v^2) \Delta t_{Bob}^2$$

Plugging-in our time-dilation result for Bob's time interval in terms of Ann's, we get:

$$\Delta s_{Bob}^2 = (c^2 - v^2) \left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \Delta t_{Ann} \right)^2 = c^2 \Delta t_{Ann}^2$$

So the spacetime interval measured by both Bob and Ann are the same.

This page titled [2.2: The Nature of Time](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [Current page](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).
- [2.1: Spacetime Diagrams](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

2.3: More Thought Experiments

Doppler Effect

We know that two observers in motion relative to each other measure the same speed when they look at a light wave, but what about the other properties of the light, such as the frequency and wavelength? We know that for sound they would not measure the same frequency due to the *doppler effect* (the phenomenon responsible for the change in perceived pitch of a car siren as it drives by), but in the case of the doppler effect for sound the medium through which the sound travels plays a critical role. In particular, an observer moving through the air toward a sound source will note that the sound wave is moving toward them faster than that sound is moving through the air. This of course is not the case for light traveling through a vacuum. Nevertheless, there is a doppler effect for light. To get started, we need to define a few things about waves:

The length of the repeating waveform, called the *wavelength* of the wave, we represent with the symbol λ . A snapshot of the wave tells us something about its spatial features like the wavelength and amplitude, 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 (or in the case of light, where no medium is needed, the strength of something called the electromagnetic field – but more on this in classes yet to come). 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 the process repeats, 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 period and frequency:

$$f = \frac{1}{T} \quad (2.3.1)$$

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):

$$c = \frac{\lambda}{T} = \lambda f \quad (2.3.2)$$

In the analysis to come, we will represent the "crests" of light waves with circles, so that the distance between these circles is the wavelength. We'll start with the basic phenomenon of doppler effect. The two gifs that follow apply equally to light or sound. The main idea is to note that when there is no relative motion, the rate of flashes of the red source equals the rate of flashes of the blue receiver (each flash of the receiver occurs when a crest arrives). But when the source is moving relative to the receiver, the rate of source and receiver flashes do not match. Specifically, the receiver frequency goes up when the relative motion is toward each other, and goes down when it is away from each other.

Figure 2.3.1 – No Relative Motion of Source and Receiver

.

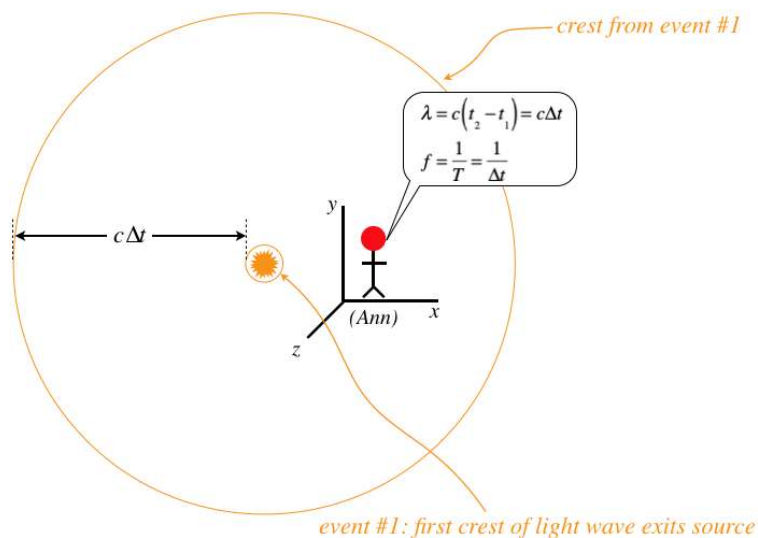
.

Figure 2.3.2 – Approaching Relative Motion (Receiver's Perspective)



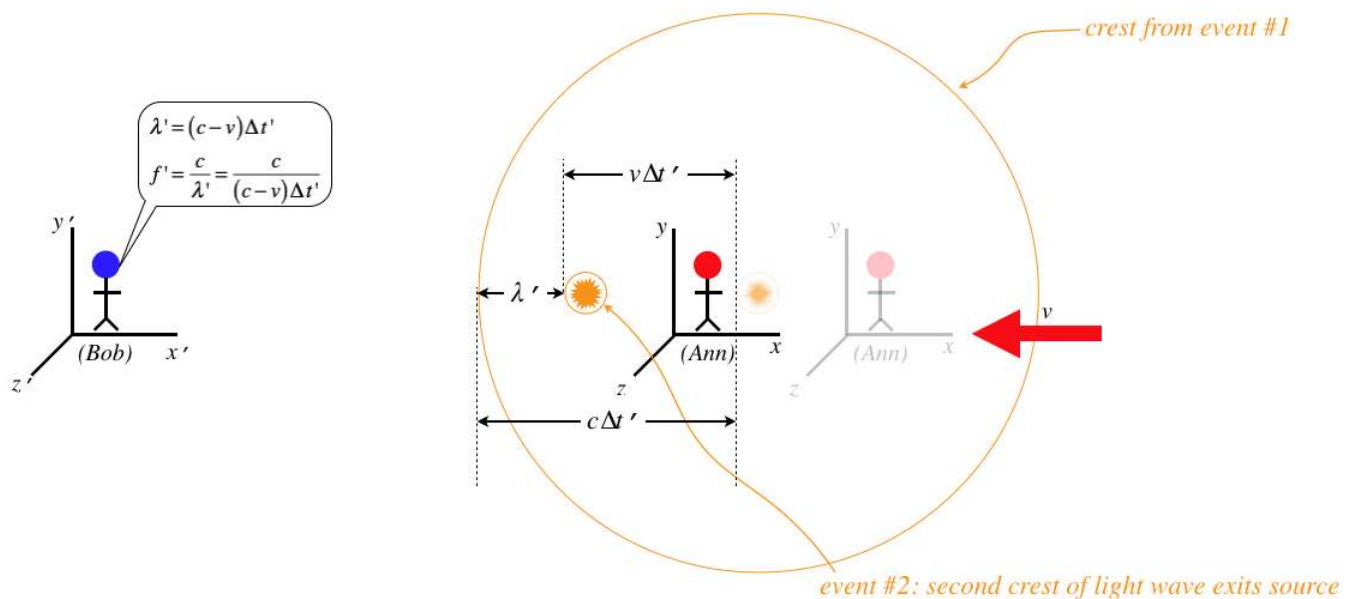
Now let's put Ann and Bob into the roles of source and receiver, respectively. We start with what Ann measures for the light source that remains stationary in her frame:

Figure 2.3.3 – Ann's Perspective of Light Signal for Two Crests



The wavelength's relationship to the frequency and wave speed is as we stated above.

Figure 2.3.4 – Bob's Perspective of Light Signal for Two Crests



Bob's measurement of the wavelength is different from Ann's because in the time between emissions of the two wave fronts, the source moves according to Bob. But this is not the only source of the disparity between the two frequency measurements. The diagram shows that the time elapsed between wave front emissions also plays a role, and as we know, the fact that the flashes occur at the same place for Ann means that her coordinate time measurement happens to equal the proper time measurement (and she is in an inertial frame, so this is also the spacetime interval), while Bob's coordinate time measurements involve events at different positions. So Bob will measure a longer time between flashes than Ann. While all waves (most notably sound) exhibit the doppler effect, the result is different for light, giving us an "ordinary" doppler effect, and a relativistic doppler effect. Plugging in the time dilation relation gives the relation between the two frequencies measured:

$$f' = \frac{c}{(c-v)\Delta t'} = \left(\frac{c}{c-v}\right) \left(\frac{1}{\gamma_v \Delta t}\right) = \left(\frac{c}{c-v}\right) \left(\sqrt{1 - \frac{v^2}{c^2}} \frac{1}{\Delta t}\right) \quad (2.3.3)$$

$$= \sqrt{\frac{c+v}{c-v}} f(\text{moving toward each other})$$

Whenever this occurs with light in the visible spectrum, the change in frequency goes away from the red end of the spectrum, and toward the blue end, so this increase in frequency for light is called a **blue shift** (even when the light is not in the visible spectrum).

If Ann happens to be moving *away* from Bob, then it is a simple change to this equation to get the correct answer – change v to $-v$, giving:

$$f' = \sqrt{\frac{c-v}{c+v}} f(\text{moving away from each other}) \quad (2.3.4)$$

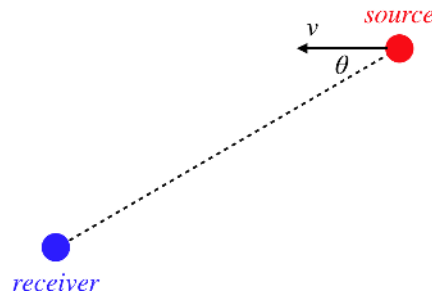
This effect of reducing the frequency perceived due to relative motion is called a **red shift**.

But of course these are not the only two options. For example, Ann and Bob may be in the process of moving past each other. If this is happening, then determining the wavelength measured by Bob is a tougher, but the time dilation between the two frames still applies. One specific example we can look at is when Bob and Ann are aligned along the y -axis. We have to be careful about using words like "when" in the context of relativity, so we will define this moment for Bob as when he *sees* the source of the light as being aligned with him along the y -axis (of course, he will *deduce* that the light source is elsewhere, but that is not what counts here, as Bob is actually observing the light). At this moment, Ann and Bob will agree upon the wavelength, since their relative motion is along the x -direction and has no effect on the spacing of wave fronts. In this case, *only* the time dilation plays a role, and the result is simply:

$$f' = \frac{1}{\Delta t'} = \frac{1}{\gamma_v \Delta t} = \frac{\sqrt{1 - \frac{v^2}{c^2}}}{\Delta t} = \sqrt{1 - \frac{v^2}{c^2}} f \quad (2.3.5)$$

The more general case involves the line joining the source and receiver forming an angle θ with the relative velocity vector:

Figure 2.3.5 – Relative Motion of Source and Receiver Not Along Line Joining Them



Once again, the line joins the receiver and the apparent (not deduced) source of the light. In this case, the doppler effect on the frequency comes from the component of the source's motion relative to the receiver that lies along the line joining them. That is we replace the v above with $v \cos \theta$:

$$f' = \frac{c}{(c - v \cos \theta) \Delta t'} = \left(\frac{\sqrt{c^2 - v^2}}{c - v \cos \theta} \right) f \quad (2.3.6)$$

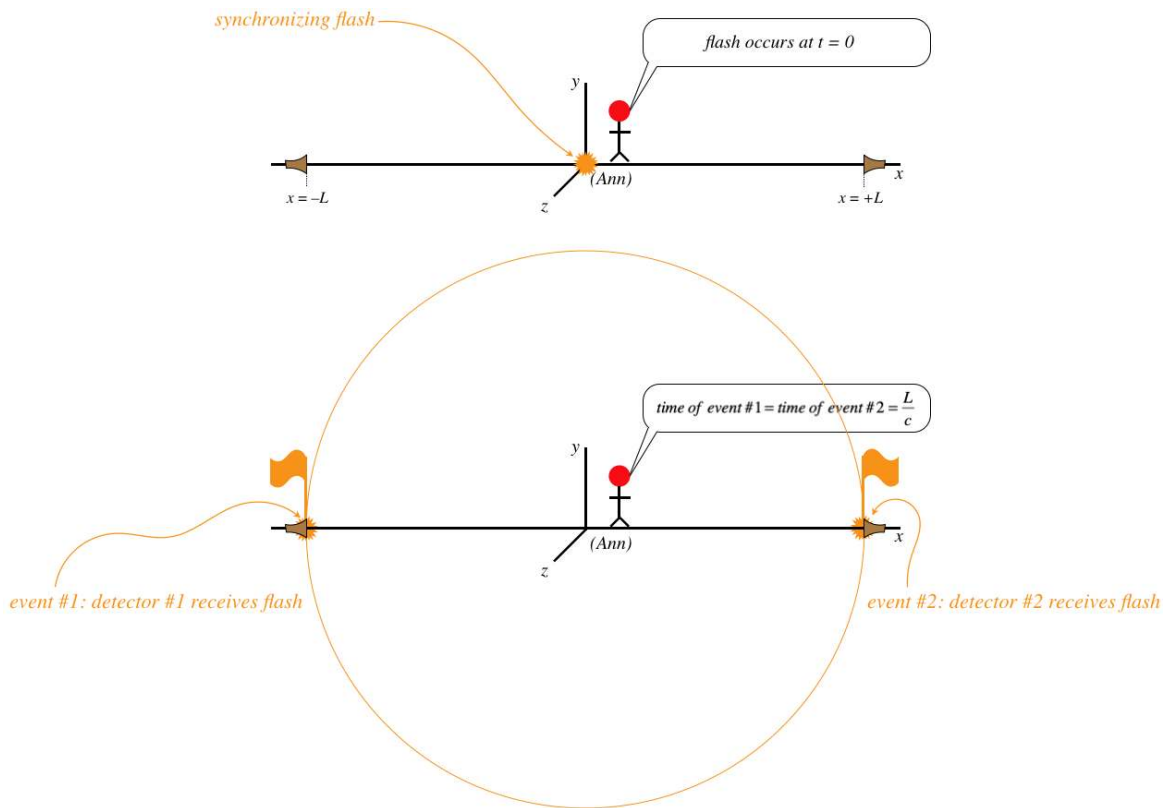
It is left as an exercise for the reader to show that this equation reduces to the three equations above for the appropriate values of θ .

Simultaneity

Let's return to our discussion of how to measure coordinate time by synchronizing clocks at all the lattice points in a reference frame. Suppose Ann and Bob are moving past each other along the x -axis, and at the moment that their origins coincide, they start their clocks at the origin. Then each of them synchronizes all the clocks on their lattice with the clock at the origin. Doesn't this mean that all of Ann's clocks are synchronized with all of Bob's clocks? And if so, doesn't this mean that they should measure the same coordinate times between events, in contradiction to everything we have said so far? Such a conundrum calls for a thought experiment!

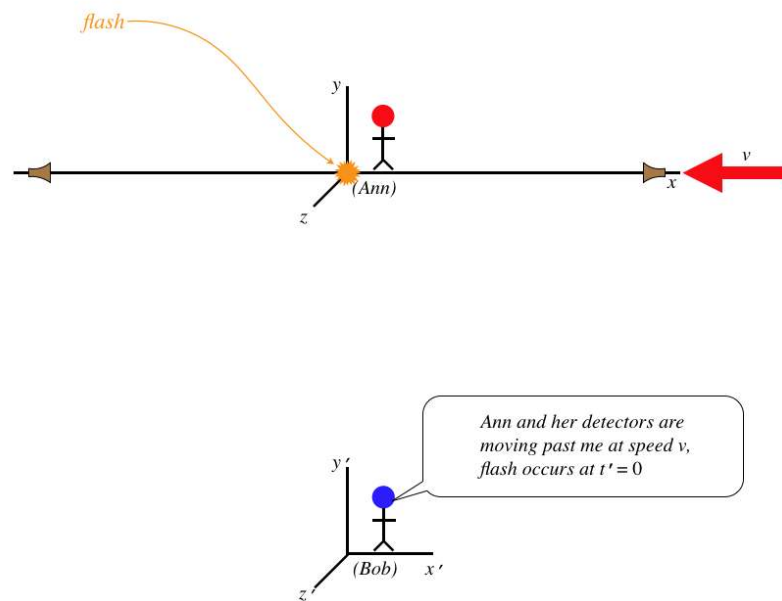
Let's suppose Ann decides to synchronize two clocks using a flash from her clock at the origin:

Figure 2.3.6 – Simultaneous Events for Ann



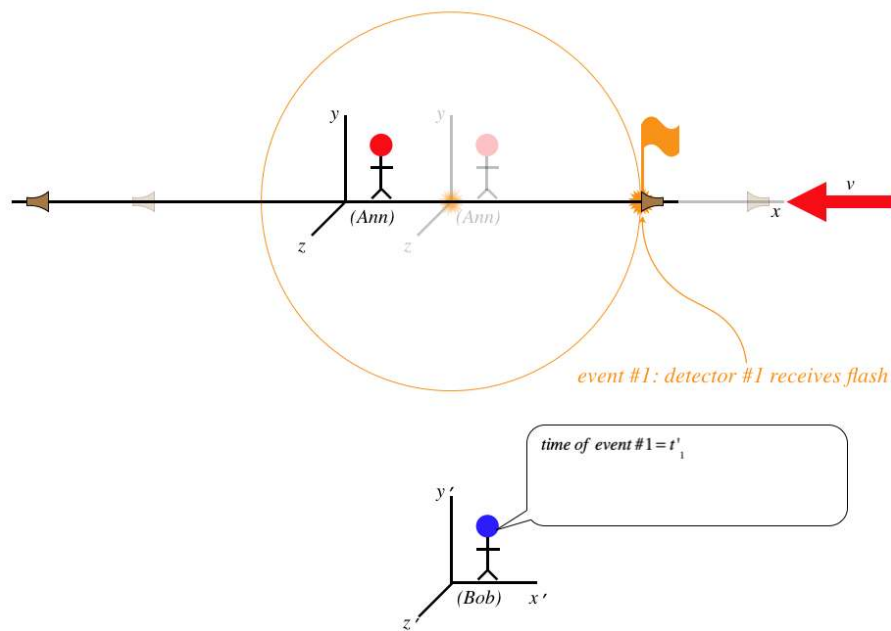
Does Bob agree that the two clocks (located at the two detectors) are synchronized? Let's look at what Bob sees:

Figure 2.3.7 – Ann's Synchronized Events Seen by Bob (a).



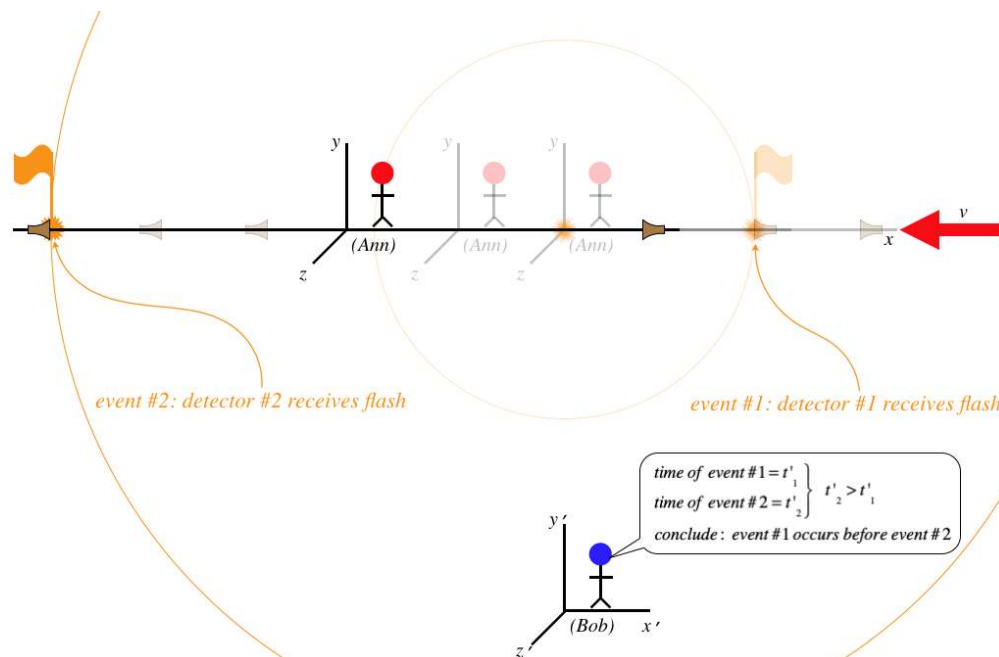
The detectors Ann is using are fixed in the lattice points in her frame, so they move along with her, according to Bob. When the light flashes, it takes time for the wave to get to the detectors, and while this time passes, the detectors move, according to Bob:

Figure 2.3.8 – Ann's Synchronized Events Seen by Bob (b).



As you can see, the detector trailing Ann receives the signal before the other detector, according to Bob.

Figure 2.3.9 – Ann's Synchronized Events Seen by Bob (c)



Far from seeing the two events simultaneously, Bob measures a time difference between them. This means that when he looks at all the clocks in Ann's lattice, he sees them all out of sync, with the times getting later the farther the clock is on the positive side of the origin. We therefore find that the concept of simultaneous events is relative (frame-dependent).

Alert

It is important to keep in mind that when we are talking simultaneous events in one frame, we are not talking about simply seeing two things occur at a different time. For example, if Ann happened to be standing close to detector #1, then the light from the flag that pops up there would reach her sooner than the light coming from the flag at detector #2, and she would witness the two flags popping at different times, but the two events would still be simultaneous in her frame.

Example 2.3.1

We found in the light clock thought experiment that the relationship between Ann's and Bob's time measurements is given by Equation 1.2.2. If Ann's two clocks are synchronized, then the time between the two events that occur when the flash reaches both detectors is zero. So why don't we find that for those same two events viewed by Bob, the time interval is also zero?

$$\Delta t' = \gamma_v \Delta t = 0$$

Solution

The equation quoted assumed that the time measured by Ann was the proper time, since the two events occurred at the same position in her inertial frame. The synchronized events in this case do not occur at the same position, so it is the coordinate time that she measures to be zero. One way to avoid this confusion is to write the time dilation formula of Equation 1.2.2 explicitly in terms of the proper time:

$$\Delta t' = \gamma_v \Delta \tau$$

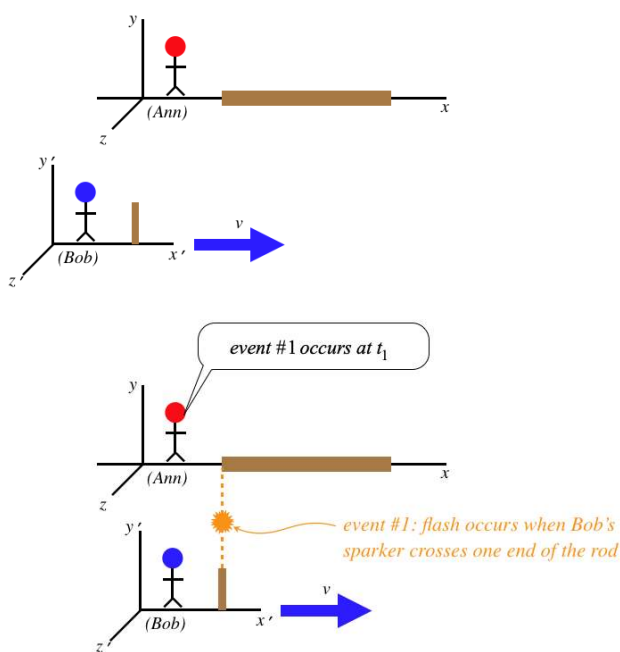
In the case above, the comparison is not between a coordinate time and a proper time, but two different coordinate times.

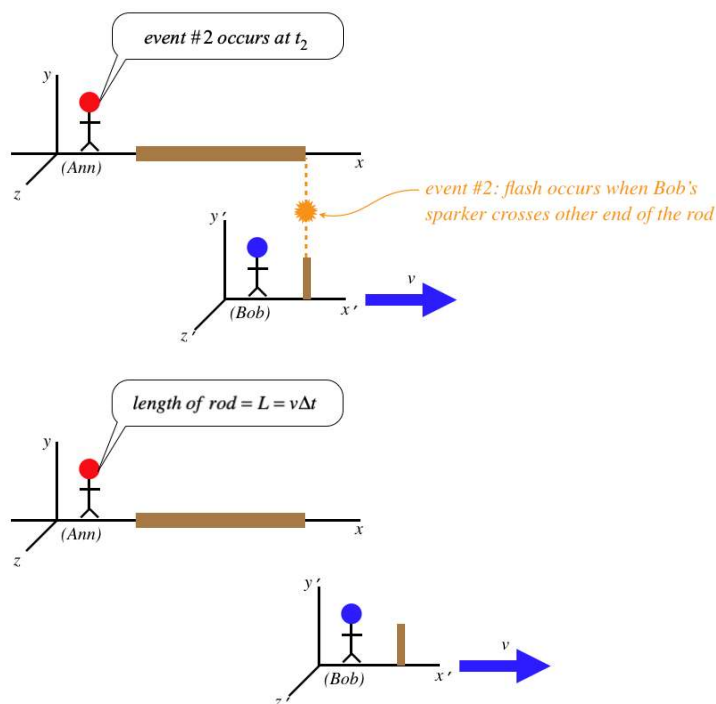
Length Contraction

Instead of comparing time spans between two frames in relative motion, let's compare distance spans. To do this, we need to first figure out *what it means* to measure the length of an object (say a meter stick). As we know, whatever we do in relativity must be in terms of spacetime events. We can't simply say that the length of an object is the distance between events that occur at the object's endpoints, because the object might move after one event occurs and before the second one occurs. So clearly to define the length of an object, we need to stipulate that the two events that occur at the endpoints of the object being measured occur *at the same time*. But since observers in two frames in relative motion will not agree to what events are simultaneous, it stands to reason that they might not agree to length measurements.

We consider The following scenario: Ann lays a rod down along her x -axis, and as she passes by Bob, each end of the rod creates a spark (constituting a spacetime event) when it coincides with a device stationary in Bob's frame that we will call a "sparker."

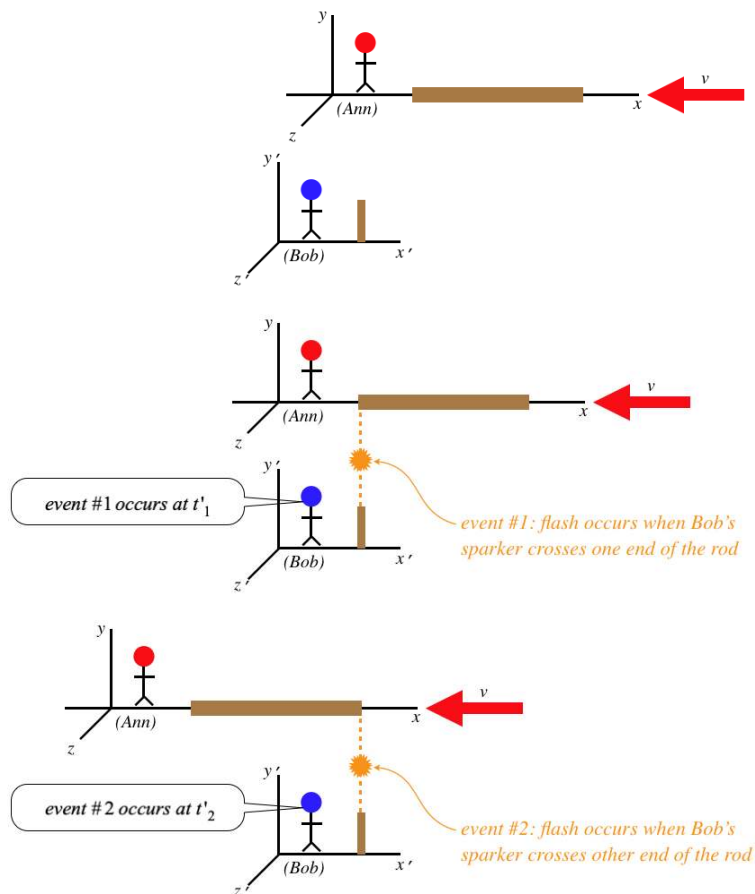
Figure 2.3.10 – Ann Measures the Length of the Rod

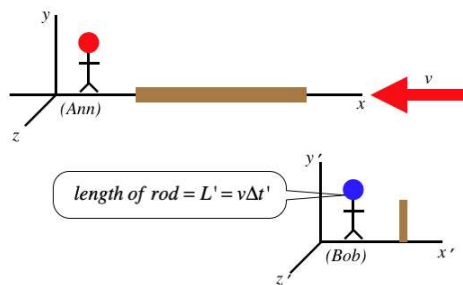




Ann measures the length of her rod to be the speed of Bob's sparker multiplied by the coordinate time she measures between the two events. The way Bob measures the length of the rod is similar:

Figure 2.3.11 – Bob Measures the Length of the Rod





We can now use the time dilation formula that relates these two times to determine a relationship between the two measured lengths, *but we have to be very careful here*. Namely, we must ask ourselves, whose measures the dilated time here? Put another way, which of these two observers measures the proper time between the two events? The answer is clearly Bob, since both events occur at the end of his sparker, which is at rest in his frame, meaning that both events occur at the same place – the exact criterion for proper time. Therefore we find that it is Ann's time between events that is longer than the time measured by Bob, giving:

$$\Delta t = \gamma_v \Delta t' \Rightarrow L' = v \Delta t' = v \frac{\Delta t}{\gamma_v} = \frac{L}{\gamma_v} \quad (2.3.7)$$

So Bob measures the rod to be *shorter* than Ann measures it to be (recall $\gamma_v > 1$). This phenomenon is known as *length contraction*.

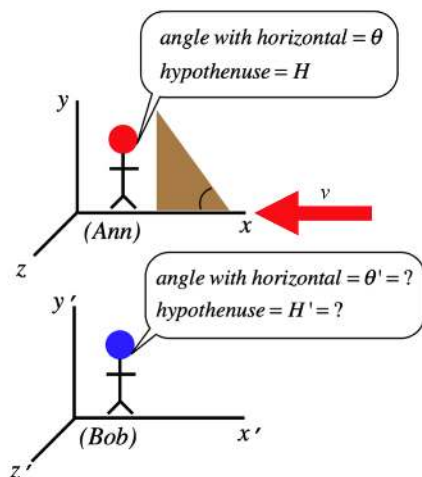
A few comments about this result:

- The longest possible measurement of length occurs in the rest frame of the object whose length is being measured. This length is often referred to as the *proper length*.
- We are accustomed to attributing different visual observations of the same object to optical illusions. This is *not* one of those cases. The same rod is *shorter* for Bob than it is for Ann. Length is not an intrinsic property – it is observer-dependent.
- If Bob zooms by Ann with an identical rod, then Bob will measure Ann's rod to be shorter than his own, *and* Ann will measure Bob's rod to be shorter than hers. We will explore this seeming paradox soon, but the short answer is that *length is not a quality that is inherent to an object*, so the fact that the rods are identical does not mean that their lengths are. Most people are not bothered by the fact that two identical rods may have different colors (due to red/blue shift), because it isn't too difficult to accept that color is not a property inherent to objects, but length is significantly tougher to swallow. If it helps, it's okay to say that identical objects have equal proper lengths.
- If we consider examining the same rod in both frames when it is aligned along the y -axis (while relative motion is still along the x -axis), we will fairly easily find that both observers agree on the length. Put another way, *lengths contract only along the dimension parallel to the direction of relative motion*.
- In terms of the lattice points for the two frames, we would say that Bob sees Ann's lattice points are closer together along the x -direction than his own, and Ann would conclude the same about Bob's lattice points. This is because each can measure the rod in terms of these points, so if the length of the rod is relative, so is the entire space of the moving item.

Angles

Consider our usual setup with Ann and Bob in relative motion along their common x -axis. We know that there is a contraction of length along the x -axis when an object is moving relative to the frame it is viewed in. Furthermore, we know that lengths along the y and z axes do not contract. Suppose Ann has a right-triangular wooden block at rest in her frame as in the figure below, and measures the angle it makes with the x -axis to be θ . What angle does Bob measure?

Figure 2.3.12 – Ann and Bob Measure an Angle



With only the side of the triangle along the x -axis contracting, the angle must change. If we call the length of the base of the triangle in Ann's frame x and the height y , then we get:

$$\theta' = \tan^{-1} \left(\frac{y'}{x'} \right) = \tan^{-1} \left(\frac{y}{\frac{x}{\gamma_v}} \right) = \tan^{-1} \left(\gamma_v \frac{y}{x} \right) = \tan^{-1} (\gamma_v \tan \theta) \quad (2.3.8)$$

Example 2.3.1

Compute Bob's unanswered question in the figure above – what does he measure for the hypotenuse of the triangle, in terms of the hypotenuse and angle measured by Ann?

Solution

Start with the Pythagorean theorem:

$$H' = \sqrt{x'^2 + y'^2} = \sqrt{\frac{x^2}{\gamma_v^2} + y^2} = \sqrt{\left(1 - \frac{v^2}{c^2}\right) x^2 + y^2} = \sqrt{x^2 + y^2 - \frac{v^2}{c^2} x^2}$$

Now write Ann's values of x and y in terms of H and θ :

$$\left. \begin{array}{l} H^2 = x^2 + y^2 \\ x = H \cos \theta \end{array} \right\} H' = \sqrt{H^2 - \frac{v^2}{c^2} H^2 \cos^2 \theta} = H \sqrt{1 - \frac{v^2}{c^2} \cos^2 \theta}$$

This page titled [2.3: More Thought Experiments](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [Current page](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).
- [2.2: The Nature of Time](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

2.4: Paradoxes

The Ladder & Barn Paradox

We have derived some rather outlandish results from the postulate(s) of relativity, using the simple but powerful tool of thought experiments. We will now use that same tool to construct some scenarios that appear at first blush to result in logical inconsistencies in the theory called *paradoxes*, and then we will shoot them down, demonstrating that the theory is in fact logical and consistent.

The first paradox we will examine involves length contraction, and it goes like this...

A farmer wishes to store a long ladder that he owns inside his barn, but is frustrated to discover that his ladder is too long to fit. Specifically, he finds that his ladder is 50ft long, while his barn is only 40ft long. But like every good farmer, this fellow is well-versed in special relativity, and decides that if he can get the ladder moving fast enough relative to the barn, then the ladder's length will contract enough so that both the front and back barn doors can be closed at the same time while the ladder is inside. Before he actually tries to close the ladder in, he does a test run, zooming the ladder in one door and out the other. Sure enough, as his wife drives their souped-up tractor at $0.6c$, he notes that both ends of the length-contracted ladder are briefly within the confines of the two doors:

$$\text{farmer measures ladder length to be: } L' = \frac{L}{\gamma_v} = \sqrt{1 - \left(\frac{0.6c}{c}\right)^2} (50 \text{ ft}) = 40 \text{ ft} \quad (2.4.1)$$

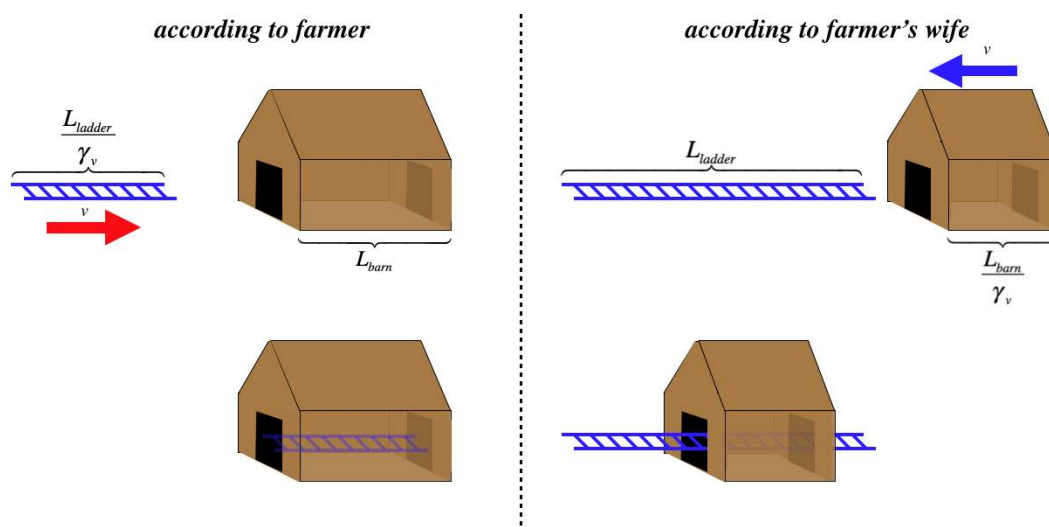
Since the 40ft barn is at rest relative to the farmer, the 40ft length-contracted ladder just barely fits.

Overjoyed that his ploy is going to work, he goes to embrace his wife, who looks distraught. When he asks her what is the matter, she says, "When I am zooming into the barn, it is way too short to fit the ladder." Indeed:

$$\text{farmer's wife measures barn length to be: } L' = \frac{L}{\gamma_v} = \sqrt{1 - \left(\frac{0.6c}{c}\right)^2} (40 \text{ ft}) = 32 \text{ ft} \quad (2.4.2)$$

The frame of the farmer's wife sees the ladder at rest and the barn moving past, so naturally the 50ft ladder doesn't fit inside the 32ft length-contracted barn. So how is it possible that both the farmer and his wife are correct at the same time? How can the ladder both be contained in the barn and not contained at the same time?

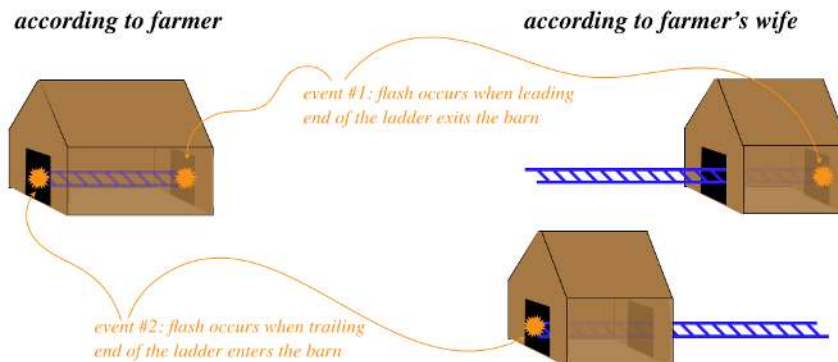
Figure 1.4.1 – Ladder & Barn Paradox



As with everything else in relativity, we can't trust our logic without converting everything into the language of spacetime events. In terms of events, what does it mean for the ladder to be "entirely within the barn?" Well, imagine that each end of the ladder is equipped with a flashbulb that flashes whenever it is in a doorway of the barn. We can say that the ladder is just barely "completely enclosed" if the light at the front of the ladder flashes in the exit doorway at the same instant that the light at the rear of the ladder flashes in the entrance doorway. Then we can say that both ends of the ladder are inside the barn at the same time. But we know

from our discussion of simultaneity that what one observer sees as simultaneous events will not be seen as simultaneous by another. So the farmer sees the ladder as being within the barn because he can declare both ends to be within the confines of the barn *at the same time*, even as his wife claims that the front of the ladder exits the barn well in advance of the rear of the ladder entering it.

Figure 1.4.2 – Paradox Resolved By Discarding Simultaneity



Once we accept that the idea of simultaneity is not universal, we realize that "being inside the barn" is a relative concept – two observers don't need to agree on this.

Okay, so much for the "dry run" of driving the ladder into one door and out the other. If the farmer can correctly claim that the ladder is within the barn, what happens if he closes the two barn doors simultaneously, while the ladder is in there? What will his wife see then?

If the barn door closings are simultaneous to the farmer, then they are not so for the wife. She would see the exiting barn door closed first, just as the front of the ladder gets there, and at that moment in her frame, the rear of the ladder will not be in the barn, and the entrance door is not yet being closed. But the question is, if the front of the ladder is stopped by the closed door, how does the rear of the ladder ever get in?

The answer to this is subtle and a bit beyond what we have discussed so far. Before we go into it, it is necessary to point out that everything we have been discussing to now points to a "cosmic speed limit" that is the speed of light. If two frames could move at the speed of light relative to each other, then one would witness time in the other frame to be stopped, and lengths to be shrunk to zero parallel to the direction of motion. This is an asymptotic speed limit that applies not only to solid objects, but also *influences* (such as forces and fields) and *information*.

Given that influences can't travel faster than c , we must concede that even if we could somehow instantly stop the front of the ladder with the closed barn door, the *rear of the ladder will keep moving forward* until it "gets the message from the front" to stop. From the wife's perspective, this means that the front of the barn will continue moving at v until the rear of the barn sends a message to it to stop moving. As time passes after the collision of the back door of the barn with the ladder, more and more of the barn will stop moving, until finally the front of the barn gets the message and the entire barn is at rest. If the message to stop could get to the front of the barn instantly, then of course the front door could never be closed and we would have a paradox, since the farmer clearly sees that it is possible to close both doors with the ladder within. So we need to calculate to see if the message is slow enough (even at the cosmic speed limit) to stop the front of the barn before the ladder can be enclosed.

Figure 1.4.3 – Can Both Barn Doors Be Closed?

occurs during both the departure and the return, we see that both twins measure the proper time between these events, because the positions of both events in space are the same.

But is this symmetric? Both twins start in the same inertial frame, but in order to end up in different inertial frames, one of them has to *change frames*. For this to happen, she needs to accelerate, and this (at least while she is accelerating) puts her into a special non-inertial frame, which she *can* do an experiment to detect. The flash of light at the space dock occurs at the same position in each twin's frame, so they both measure a proper time, but only Ann measures the spacetime interval between those flashes, as only she stays in the inertial frame of the clock that records both events.

This means that the viewpoint of Ann remains correct – her sister experiences fewer years than she does during the trip. They started as twins, but after the trip Ann is several years older.

Alert

It is important to understand that this is not the fountain of youth – both twins age in the usual way. It isn't the rate of aging that is changed, it is the rate of time flow itself that is different. The older sister has more experiences than the younger one, and if they both live to be 100, they have equally-long lives. Their lives just don't conclude at the same time, even though they started at the same time.

Let's do a specific calculation of this effect. Suppose that before the trip begins, Bea tells Ann that she's going to a destination that is 20 light-years away and coming back. Her ship will travel at a more-or-less constant speed of $0.8c$ (except for the quick accelerations near the two endpoints). Ann does the simple math and determines that to cover the distance of the round trip at that speed will require 50 years:

$$\Delta x = v\Delta t \Rightarrow \Delta t = \frac{\Delta x}{v} = \frac{40 \text{ light years}}{0.8c} = 50 \text{ years} \quad (2.4.6)$$

And this is in fact the amount of time that Ann measures for Bea's trip. But what does Bea measure? Once Bea quickly gets up to speed, the distance isn't so far anymore, as the separation of the two endpoints is length-contracted (think of placing a very long rod between the endpoints). What is 20 light-years away in Ann's frame in Bea's frame is:

$$L' = \frac{20 \text{ light years}}{\gamma_v} = \sqrt{1 - \left(\frac{0.8c}{c}\right)^2} (20 \text{ light years}) = 12 \text{ light years} \quad (2.4.7)$$

At a speed of $0.8c$, this 24 light-year round trip only takes 30 years. So Bea returns home to find that her twin sister is 20 years older than she is.

Let's call the departure of Bea and her arrival spacetime events A and B. Note that these both occur at the same place in space (Ann & Bea's home), which means that *both* Ann and Bea measure a proper time between the events. The fact that these are different comes back to [important note number 2](#) regarding proper time, which states that each proper time measurement is unique to the history of the clock that measures it. Ann remains in an inertial frame, but Bea must accelerate to leave, turn around, and stop when arriving home. Just stating this is so is unsatisfying, so let's try a new thought experiment to clear it up. This involves three players: Ann, Bob, and Chu, who have the following roles:

- Ann will be our “reference” observer. We will watch the motions of the other two through a telescope from the comfort of her (by her account, stationary) inertial frame on her spaceship.
- Bob will be engaging in a race from the starting point to the finish in a spaceship that travels in a straight line at a constant rate (as measured by Ann) of $0.5c$. So Bob remains in an inertial frame throughout the trip.
- Chu will be the other contestant in the race, but he will get off to a slow start with a speed of $0.25c$. Sometime during the race he will instantaneously speed his spaceship up to $0.75c$ (both speeds according to Ann) in an effort to catch up to Bob before the finish line.

All three participants have identical clocks on board their ships, and when the race begins, a flash of light (spacetime event) at the start line signifies the start of the race. When a ship crosses the finish line, another flash of light is given off there, providing a spacetime event that indicates that the ship has finished the race. Ann measures the distance between the start and finish lines as being 40 light-minutes (roughly the distance from our Sun to Jupiter).

Let's compute the time it takes Bob to complete the race, according to Ann. This is easy - no relativity necessary. Traveling at $0.5c$ over a distance of 40 light-minutes requires 80 minutes.

According to Ann, Chu makes his shift in speeds 40 minutes into the race (when Bob is halfway to the destination). So in the first 40 minutes at a speed of $0.25c$, Chu travels 10 light-minutes. Then for the remaining 30 light-minutes of the trip he is traveling at $0.75c$, which means it takes him 40 more minutes... He finishes in a tie with Bob!

The start and finish of the race both occur at the same position in Bob's frame (the front tip of his ship), and he is in an inertial frame, so he measures the proper time interval $\gamma_{0.5c} \Delta\tau_{Bob} = \Delta t_{Ann}$, as we know from the time dilation formula between two inertial frames. Bob's time comes out to be:

$$\Delta\tau_{Bob} = \frac{\Delta t_{Ann}}{\gamma_v} = \frac{\sqrt{3}}{2} (80min) \approx 69min \quad (2.4.8)$$

Now let's compute the time measured by Chu. We can do this by splitting his trip into three events. We already have two of them – the flash that occurs at the start and finish. As with Bob, these two events occur at the same place in his frame, so he measures a proper time, but as his inertial frame changes, it is not the same. The third event we will define as a flash at the front tip of Chu's spaceship when he suddenly changes speeds. Now we can compute his time for each leg of the trip separately. He remains in an inertial frame during each leg, so we can use the time dilation of Ann for each leg (like we did for Bob's whole trip):

$$\begin{aligned} \Delta\tau_{Chu} &= \Delta\tau_{Chu}(leg\ 1) + \Delta\tau_{Chu}(leg\ 2) = \frac{\Delta t_{Ann}(leg\ 1)}{\gamma_{v_1}} + \frac{\Delta t_{Ann}(leg\ 2)}{\gamma_{v_2}} \\ &= \sqrt{1 - 0.25^2} (40min) + \sqrt{1 - 0.75^2} (40min) = \frac{\sqrt{15} + \sqrt{7}}{4} (40min) \approx 65min \end{aligned} \quad (2.4.9)$$

So we see that less time elapses for Chu than for Bob during this race. So how does this apply to the twin paradox? Let's view the whole race from Bob's perspective. Let's call the direction of the race the $+x$ -direction. According to Bob, he and Chu start at the same position, and Chu is initially going in the $-x$ -direction (because Bob is moving faster than Chu in the $+x$ direction according to Ann). A little while later, Bob sees Chu suddenly stop and immediately start coming back toward him (because Chu is now moving faster than Bob in the $+x$ direction according to Ann), until they are back at the same position. Bob insists he was stationary the whole time, so according to Bob, Chu basically took off and came back. And sure enough, when he came back, Chu had aged 4 minutes less than Bob.

While we haven't proven it with this one example, we do see a general result: For the same two spacetime events, all proper time measurements (like Bob's and Chu's) are shorter than all coordinate time measurements that are not proper (like Ann's). Also, the proper time interval between two events measured in an inertial frame (like Bob's) is *longer* than all of the other (non-inertial) proper time measurements (like Chu's).

This page titled [2.4: Paradoxes](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

CHAPTER OVERVIEW

3: Kinematics in Special Relativity

[3.1: Spacetime Diagrams](#)

[3.2: Lorentz Transformation](#)

[3.3: Velocity Addition](#)

[3.4: Electricity and Magnetism](#)

This page titled [3: Kinematics in Special Relativity](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

3.1: Spacetime Diagrams

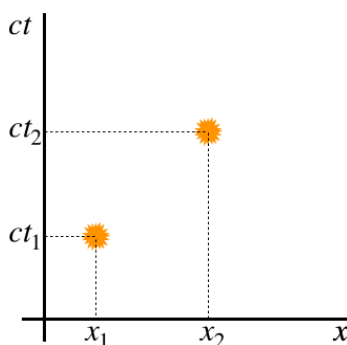
Graphing Spacetime Events

Up to now, we have been representing the position of a spacetime event relative to the spatial axes of Ann and Bob, while representing time by drawing the axes at different positions, with the set of axes remaining fixed being the coordinate system of the observer from whose perspective we are seeing the event. Here we will represent spacetime events in a more efficient and useful (though perhaps a bit more abstract) manner.

We will stay with simplified situations where all the action occurs along the x -axis, which means that we have room for something else. Specifically, we will draw the position (along x -axis), and *time* axes for an observer in an inertial frame whose perspective we are viewing from. With this construction, a spacetime event is simply a point located in the plane of the two axes, and the coordinate position and coordinate time for the observer we are viewing from can be read off the axes directly.

In 9A we worked with 1-dimensional graphs with the position represented on the vertical axis, and the time on the horizontal axis. In relativity, it is standard (perhaps just to alert the reader to the fact that relativity is being considered) to use time as the vertical axis. What is more, these diagrams give both axes the same units by scaling the vertical axis by the speed of light, c . The resulting representation of spacetime events is called a *spacetime or Minkowski diagram*.

Figure 3.1.1 – Spacetime Events on a Minkowski Diagram



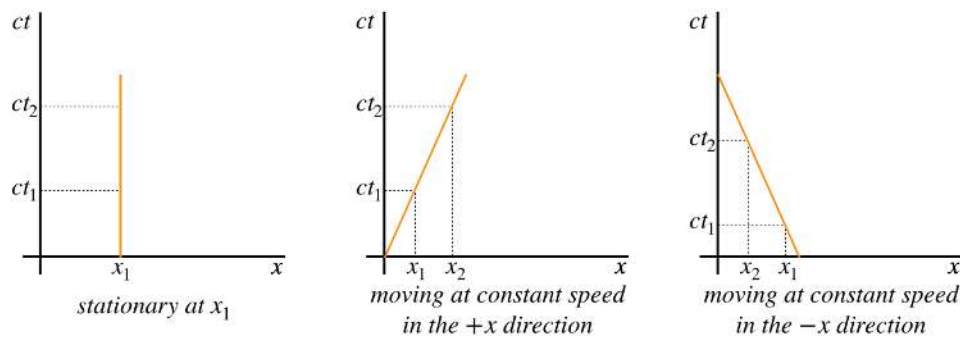
The times indicated by the values on the vertical axis are those measured by synchronized clocks in the frame of the observer represented by these axes. If the two events happen to be aligned vertically, then the two events occur at the same position in this frame, which means the difference $t_2 - t_1$ in that case is also the proper time interval $\Delta\tau$. The spacetime interval is related to the "distance" separating these events in the diagram. If the spacetime interval is measured in an inertial frame, then it is the "length" of the straight line connecting the events, otherwise it is the "arclength" of the curve connecting the events. All of these quoted words must be very confusing at the moment, but things will be made clearer below.

World Lines

While it is nice to be able to draw two separate events on the same diagram, rather than drawing multiple drawings as we did before, there is an even greater advantage to the tool of the Minkowski diagram – we can represent *spacetime trajectories*, also known as *world lines*. Suppose we track a flashing beacon as it moves along the x -axis. Each flash records a spacetime event with the position it is located and the time of the flash, so we get a string of several points on the spacetime diagram. If the flashing frequency is increased to infinity, then the spacetime points form a continuous curve that tracks the spacetime history of the moving object.

Let's consider a few special world lines. When someone observes a stationary object, its trajectory in the Minkowski diagram must be such that the position doesn't change (but of course the time does). If this person witnesses an object moving at a constant speed, then the world line is sloped, positively when the motion is in the $+x$ direction, and negatively when the motion is in the $-x$ direction.

Figure 3.1.2 – Some Simple World Lines



A closer look at the values of the slope of a world line reveals that it is:

$$\text{slope of a world line} = \frac{ct_2 - ct_1}{x_2 - x_1} = \frac{c}{u}, \quad (3.1.1)$$

$$u \equiv \frac{\Delta x}{\Delta t} = \text{speed of the object measured by observer}$$

That is, the inverse of the slope of the world line of an object is the fraction of the speed of light that the observer measures for that object, with the sign representing the direction of motion. If the "object" happens to be a light beam, then its slope is ± 1 .

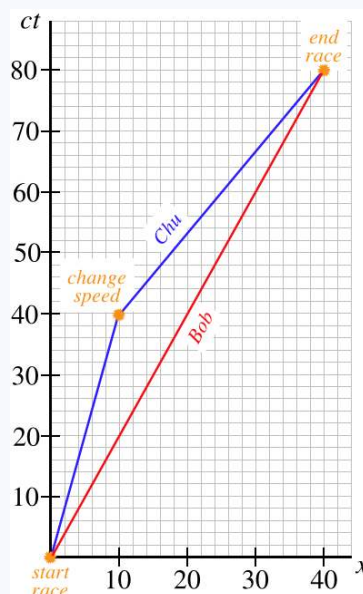
Example 3.1.1

Let's return to the thought experiment for the twin paradox involving Ann, Bob, and Chu at the end of [Section 1.4](#), and use a spacetime diagram to depict the race.

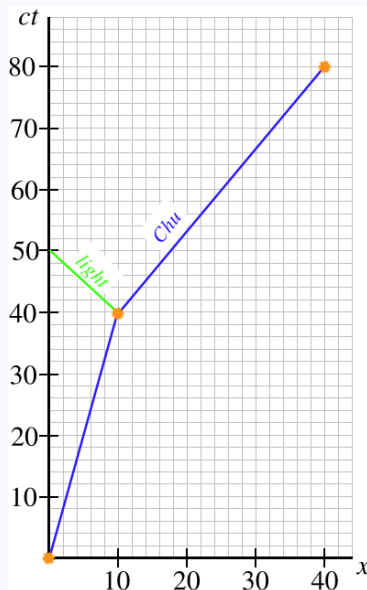
- Draw a spacetime diagram in Ann's reference frame depicting the world lines of both Bob and Chu, and label the important spacetime events along these worldlines.
- Use the diagram to determine the time on Ann's clock in her spaceship (not at the lattice point in her reference frame) when she **sees through her telescope** that Chu has changed speed.
- Use the diagram to determine the time on Bob's clock in his spaceship when he **sees through the window of his spaceship** that Chu has changed speed.

Solution

a. The diagram below has light minutes as units on both axes. Both start the race at the same point in spacetime ($x = 0$, $t = 0$ in Ann's frame), and end at the same point in spacetime ($x = 40$ light-minutes, 80 minutes later). Chu changes speed at $x = 10$ light-minutes, after $t = 40$ minutes. Bob moves at a constant speed of half the speed of light, so the slope of his world line is 2. Chu moves at one-quarter the speed of light and then three-quarters the speed of light, so the slopes of his world line segments are 4 and $\frac{4}{3}$, respectively.

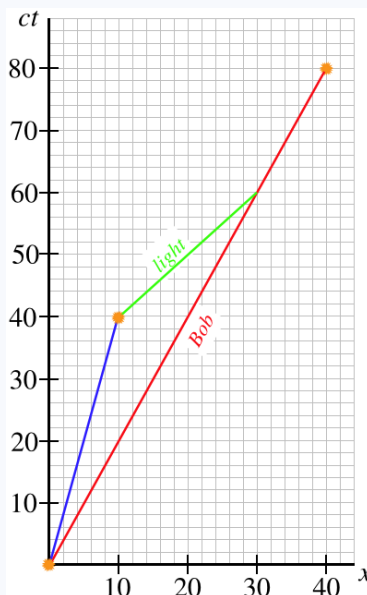


b. Ann doesn't see the event of Chu changing speed until light from that event reaches her. The world line of this light has a slope of -1 because it is coming back to Ann (in the $-x$ -direction). Drawing this into the diagram shows that she discovers Chu has changed speeds at time $t = 50$ minutes, because that's when the light world line intersects with Ann's world line (which is the vertical axis, since she starts at her origin and never moves).



c. Now we want to know when Bob sees light from the event of Chu changing speed. This time the light goes in the $+x$ -direction to get to Bob, so it has a slope of $+1$. This allows us to find the intersection point of the light world line with Bob's world line, but then deriving from that the time on Bob's clock requires a little more work. The simplest way to get this number is to find what the time is on Ann's clock, then use the time dilation formula to get Bob's time. Looking at the diagram below, we see that the time of this intersection in Ann's frame is 60 minutes, which in Bob's frame translates to:

$$t_{Bob} = \frac{t_{Ann}}{\gamma_v} = \sqrt{1 - 0.5^2} (60 \text{ minutes}) \approx 52 \text{ minutes}$$



Minkowski Spacetime

Let's represent on spacetime diagrams a comparison of what Ann sees to what Bob sees when they witness the same object's motion. Specifically, let's say they both observe the same clock, which gives off two light flashes at different times, and records those times. Let's further have the clock remain stationary in Ann's frame between these flashes.

With the clock at rest in Ann's frame, the world line for it looks like the first graph in the figure above. With Bob moving in the $+x$ -direction relative to Ann at a speed v , he sees this same object moving in the $-x$ -direction at a speed v , which means the worldline he sketches for the object looks like the third graph in the figure above.

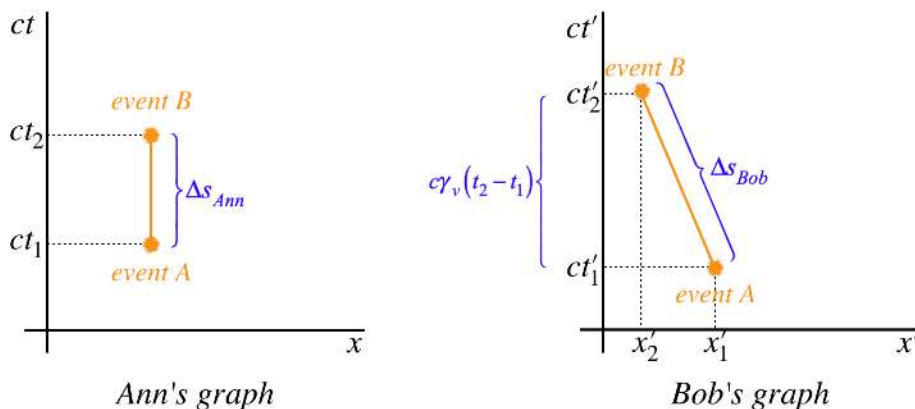
As we showed above, the spacetime interval between the two events is the *square of the length of the segment of the world line* connecting the two events in Ann's spacetime diagram. When we first discussed the nature of time, we made a point of noting that all proper times, and therefore the spacetime interval, are invariants, which means that everyone measures the same value. This would seem to indicate that if Bob measures the length of the segment of the world line he draws between the two events, he should get the same result. In fact this is true, but as we will see, there is a surprise twist.

We found that the time dilation effect gives the comparison between time intervals to be (Bob is the primed frame here):

$$t'_2 - t'_1 = \gamma_v (t_2 - t_1) \quad (3.1.2)$$

But wait, since $\gamma_v > 1$, this means that $ct'_2 - ct'_1 > ct_2 - ct_1$, and when we look at the spacetime diagrams for Ann and Bob, we see this means that the vertical change in the graph is greater for Bob than it is for Ann.

Figure 3.1.3 – Comparison of Spacetime Intervals for Ann and Bob



It sure doesn't look like there is any way that Ann and Bob could agree on the length of the spacetime interval between these two events! Now for the twist (and the explanation for why we used the quotes on the words "distance", "length", and "arclength" earlier). Let's go back to the light clock thought experiment from [Section 1.2](#). The distance traveled by the light beam was, according to Ann and Bob:

$$\begin{aligned} \text{Ann :} \quad c\Delta t &= 2L \\ \text{Bob :} \quad c\Delta t' &= \sqrt{(2L)^2 + (x'_2 - x'_1)^2} \end{aligned} \quad (3.1.3)$$

Eliminating the $2L$ from these equations gives:

$$c^2\Delta t^2 = c^2\Delta t'^2 - \Delta x'^2 \quad (3.1.4)$$

The left side of this result is the spacetime interval squared written in Ann's frame, and the right side of the equation is how it is written in Bob's frame, so they *are* equal. This looks very close to the Pythagorean theorem, with the exception that the squares that are added are not all positive. Now we can see how the lengths of the two world line segments are actually equal after all. The horizontal leg of the right triangle in Bob's graph actually contributes a *negative* amount to the length of the hypotenuse. Clearly we have redefined "length" in this context (which we no longer call length, preferring the word "interval" to avoid confusion), but it is necessary to do so to accommodate the nature of time and the constancy of the speed of light.

We therefore reiterate explicitly what we said about the invariance of the spacetime interval back in [Section 1.2](#):

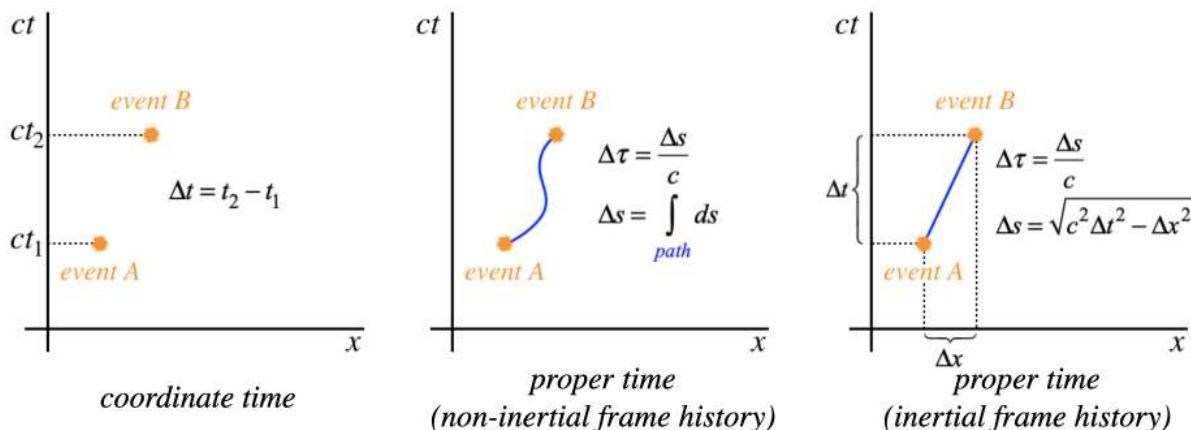
$$\Delta s'^2 = \Delta s^2 \Rightarrow c^2\Delta t'^2 - (\Delta x'^2 + \Delta y'^2 + \Delta z'^2) = c^2\Delta t^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2) \quad (3.1.5)$$

Notice this works for the case of Ann and Bob above, since $\Delta x = 0$ for Ann.

Three Measurements of Time Between Two Events

With the tool of spacetime diagrams and a means for computing intervals between events, we can gain a better understanding of the three versions of time.

Figure 3.1.4 – Three Measurements of Time in Spacetime Diagrams



The figure above shows the same two events expressed in the same frame of reference. The first graph shows that the coordinate time elapsed between the two events is simply the difference in the times measured by the synchronized clocks at the two lattice points.

The middle graph expresses a general version of proper time. The time is measured by a clock whose world line passes through both events, and it equals the path length of the clock through spacetime divided by c . This path length is found by doing a line integral to add up all the tiny contributions ds – each of which is an infinitesimal straight line segment.

$$ds^2 = c^2 dt^2 - (dx^2 + dy^2 + dz^2) \Rightarrow ds = dt \sqrt{c^2 - \left(\left[\frac{dx}{dt} \right]^2 + \left[\frac{dy}{dt} \right]^2 + \left[\frac{dz}{dt} \right]^2 \right)} \quad (3.1.6)$$

The derivatives are the components of the velocity of the object tracing the world line, measured in the frame indicated by the axes, so we can right this more compactly as:

$$ds = dt \sqrt{c^2 - (u_x^2 + u_y^2 + u_z^2)} = c dt \sqrt{1 - \frac{u^2}{c^2}} \quad (3.1.7)$$

One can now imagine computing (the square root of) the spacetime interval between two events of finite separation by "chaining together" (integrating) infinitesimal intervals:

$$\Delta s = \int_{\text{along world line}} ds = \int_{\text{along world line}} c dt \sqrt{1 - \frac{u^2}{c^2}} \quad (3.1.8)$$

Given the relationship between the proper time and the spacetime interval, we get a link between the proper time interval and the coordinate time interval between events A and B:

$$\Delta \tau = \int_{\text{along world line}} dt \sqrt{1 - \frac{u^2}{c^2}} \quad (3.1.9)$$

Notice that for the middle graph, the world line is changing slope, which means that the value of u is changing along the path (the object is accelerating), and the integral is not simple. But if the slope of the world line remains constant, like it is in the right graph above, then the square root can be removed from the integral, and the result is the usual time dilation formula.

If a different observer were to graph spacetime coordinate points for the same two events, and then watch the same clocks, the three graphs would look different. The coordinate time separating the two events would change. The world lines of the clock would look different in the other two graphs as well. The world line in the right graph would still be a straight line, but with a different slope.

The world line would be different in the middle graph as well, but would still be curved. However, the calculation of Δs in these two graphs would come out the same as they did for the original observer.

Alert

The calculation of Δs for the moving clock in the noninertial frame (middle graph) is the same when computed by any observer. The calculation of Δs for the moving clock in the inertial frame (middle graph) is the same when computed by any observer. Don't confuse this for saying they are the same as each other! The time elapsed on the clock is proportional to the path length Δs through spacetime, and the path length is different for a straight line compared to a curved one.

We found in our discussion of the twin paradox that the proper time span between two events is longer when measured using an inertial clock than when measured with a non-inertial clock (the twin in the non-inertial space ship ages less over the round trip). In other words, the value of Δs in the middle graph of Figure 2.1.4 is less than the value of Δs in the right graph (remember, the world lines connect the same two events). This seems weird in light of the old adage, "the shortest distance between two points is a straight line," but remember that this is Minkowski space with the weird version of the Pythagorean theorem. To see clearly that this is true, we can change frames to one where the two events are at the same value of x – remember that these Δs values don't change when we change frames. This will change the right graph such that the world line is a vertical straight line, while the middle graph remains curvy. Because of the $-dx^2$ contributions to the integral, any horizontal deviations from the vertical path will serve to *reduce* the magnitude of the integral. So the proper time $\Delta\tau = \frac{\Delta s}{c}$ for a straight world line between two events (i.e. measured by an inertial clock) is greater than the $\Delta\tau$ for any non-straight world line between the same two events (i.e. measured by a non-inertial clock).

Example 3.1.2

Once again returning to the thought experiment for the twin paradox involving Ann, Bob, and Chu at the end of [Section 1.4](#), use the spacetime diagram created in part (a) of [Example 3.1.1](#) above to answer the following:

- Compute the (Minkowski) length of the world line for Bob to get the time elapsed for him during the race.
- Repeat the process in part (a) for Chu.

Solution

a. The x and ct parts of Bob's world line are 40 and 80 light-minutes (lm), respectively, so the Minkowski length of his world line is:

$$\Delta s = \sqrt{80^2 - 40^2} \text{ lm} \approx 69 \text{ lm}$$

Dividing Δs by the speed of light gives the time, which means that 69 minutes elapse for Bob. This agrees with the result we got in [Section 1.4](#).

b. To compute the length of the world line for Chu, we need to compute the lengths of two separate segments and then add them together.

$$\left. \begin{array}{ll} \Delta x_1 = 10 \text{ lm}, \quad c\Delta t_1 = 40 \text{ lm} & \Rightarrow \quad \Delta s_1 = \sqrt{40^2 - 10^2} \text{ lm} \approx 38.7 \text{ lm} \\ \Delta x_2 = 30 \text{ lm}, \quad c\Delta t_2 = 40 \text{ lm} & \Rightarrow \quad \Delta s_2 = \sqrt{40^2 - 30^2} \text{ lm} \approx 26.5 \text{ lm} \end{array} \right\} \quad \Delta\tau_{Chu} = \frac{\Delta s_1 + \Delta s_2}{c} \approx 65 \text{ min}$$

Again, this is in agreement with our result before.

Causality

You may have noticed that there is a peculiarity lurking in what we have discussed in this section (we [eluded to it](#) when we first discussed the "cosmic speed limit"). In case you haven't, consider the following. We start with events at two points in spacetime, and we want to measure the spacetime interval time that elapses between these events. So we start a clock at one of the events, and move it at a constant speed to the other, recording the times of each event. We know that the clock must record this time span (omitting the Δy^2 and Δz^2 terms):

$$\Delta\tau = \frac{\Delta s}{c} = \frac{\sqrt{c^2\Delta t^2 - \Delta x^2}}{c} \quad (3.1.10)$$

Naturally two events can occur anywhere in spacetime, so the question is this: If the position in spacetime is such that $\Delta x^2 > c^2 \Delta t^2$, exactly what time does the clock measure?! Does an imaginary ' i ' suddenly appear in the digital readout?

When we say that a clock cannot be moved from one event to another to record the proper time between the events, what we are really saying is that *there is no world-line that connects them*. When we think about a spacetime diagram, this means that we can place individual events wherever we like, but we can only connect them with world lines if such a line has a slope whose absolute value is greater than 1 (a speed smaller than c). This leads to three different categories for how two spacetime events relate to each other:

time-like separation: $\Delta s^2 > 0$

This is the case we have been talking about before now, where an object can follow a world line and connect the two events. There is a well-defined proper time associated with the spacetime interval. When two events are separated in this way, we can always find an inertial frame where the events occur at the same spatial position.

light-like separation: $\Delta s^2 = 0$

In this case, the world line that connects the two events must belong to a light beam. It isn't possible to change the reference frame to one that is moving at the speed of light relative to another, so although there is a world line that can connect them, it can't belong to a clock.

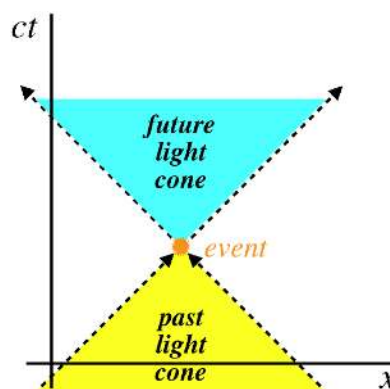
space-like separation: $\Delta s^2 < 0$

For this case, not even light can connect the two events with a world line. If we imagine an event occurring because someone at the spatial position of the event pushes a button to cause a light to flash, then the people responsible for two events separated in this way have no idea that the other event exists, as there is no world line to carry a message about one event to the other event.

This last point about space-like separated events brings up an important concept – that of *causality*. The study of physics is all about cause-and-effect – it is a net force that *causes* an acceleration, for example. We see now that we can relate two events according to whether one can cause the other. If there is no way, even in principle, for a message to get from one event to another, then there is no way for the first event to cause the other.

We can characterize what events can be causally-related nicely by using a spacetime diagram. Start by choosing a spacetime event. Then note that to be causally-related to this event, a second event must be positioned such that a straight world line drawn between the two has a slope that is no less than 1 (and if it has a slope of 1, they must be related through a signal that moves at the speed of light). This limits the region of the second event relative to the first into what is called the *light cone* of that first event.

Figure 3.1.5 – Light Cone of an Event



The light cone divides into two halves, called the *future* and *past light cones*. Only events in the past light cone can be a cause for the event at the vertex of the light cone, and the event at the vertex of the light cone can only be a cause for events within the future light cone.

It is important to keep in mind that a spacetime diagram is *based on a specific frame of reference*, and if we change to another frame, the relative positions of the events change (see Figure 3.1.3, for example). But because the spacetime interval is an invariant, what lies within the future(past) light cone of an event in one inertial frame must lie within the future(past) light cone in all inertial frames. We can put this in a far less abstract way: If in Ann's frame event #1 occurs before event #2, and the two events lie within each other's light cones, then event #1 also occurs before event #2 in Bob's frame, no matter what the relative speed of Ann and

Bob may be. This assures that if Ann declares that event #1 has caused event #2, there's no observer that can disagree with this on the grounds that event #2 has occurred first.

For a pair of events that do not lie within a light cone (such as two events that are simultaneous in one frame but at different positions, so that they are horizontally-aligned in the spacetime diagram), then the ordering of events is not necessarily preserved in all frames. To see this, let's return to the ladder-and-barn paradox from [Section 1.4](#).

Suppose the farmer's wife is driving one tractor in the $+x$ direction while the farmer's daughter is driving another tractor in the $-x$ direction in an effort to get two ladders into the barn at once (both barn doors are open, and they are coming in from opposite directions). According to the farmer, both ladders can be inside the barn at the same time, because the events at the barn doors are simultaneous in his frame. These events are at different positions, so they lie outside the light cone from each other. The farmer's wife notes that the front of her ladder reaches the back of the barn before the rear of her ladder enters, so she says that the event at the back door occurs before the event at the front door. The farmer's daughter sees the exact opposite – the event at the front door occurs first for her.

There is no way that the farmer's wife can ever claim that the event at the back door *causes* the event at the front door (even though it comes first), because her daughter would argue that it is impossible for the event that occurs second to cause the event that occurs first. Recall that the fact that a light signal cannot get from the back door event to the front door event is exactly what resolved the paradox of whether both doors can be closed with the ladder inside.

This page titled [3.1: Spacetime Diagrams](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [2.1: Spacetime Diagrams](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).
- [2.2: The Nature of Time](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

3.2: Lorentz Transformation

Transformations Between Inertial Frames

When we first studied relative motion in Physics 9HA, we wrote down a way of translating between the values measured in the two frames. This set of equations was called the [Galilean transformation equations](#). As sensible as these are, they clearly are not correct in light of what we now know about relativity. Most notably, the Galilean transformation assumes a universal time variable that is common to all frames.

So now we seek a new set of transformation equations to relate the spacetime coordinates of frames in relative motion. We will start with a couple of simplifying assumptions. First, the two frames in question share a spatial origin at the moment in time we will call $t = t' = 0$ – we will define "event A" as occurring at this spacetime point. The effect of doing this is that distances and time intervals between this event and a second event are now just the spacetime coordinates themselves. For example: $\Delta x = x - 0 = x$ and $\Delta t' = t' - 0 = t'$.

For our second assumption, we will continue to define the relative motion as the primed frame moving at a speed v in the $+x$ -direction relative to the unprimed frame.

In order to get a set of equations that gives us a translation between the (ct, x, y, z) spacetime coordinates measured in one frame and the (ct', x', y', z') spacetime coordinates measured in the other, we begin by noting that with motion only along the x -axis, the y and z coordinates will remain unchanged. For example, we know that lengths along those directions do not contract, so we would not expect the coordinates to be related in any way other than $y' = y$ and $z' = z$. But what about the x and ct coordinates?

We start by assuming that the transformation is a *linear* one, not unlike the Galilean transformation (after all, the Galilean transformation *does* work for frames whose relative speed is low). This means that the primed values can be written as linear combinations of the unprimed values:

$$\begin{aligned}x' &= J \cdot x + K \cdot ct \\ct' &= L \cdot x + M \cdot ct\end{aligned}\tag{3.2.1}$$

Our goal is to determine the unknown constants J , K , L , and M that work for relativity. Let's start by defining "event B" viewed by the primed observer. Let's say that this event occurs at this observer's time t' , and takes place at the origin of the unprimed frame. Since the primed observer sees this frame moving in the $-x'$ -direction for a time period of t' after starting at the origin, the primed observer sees this event occur at the position $x' = -vt'$. Plugging $x = 0$ (the event occurs at the unprimed origin) into the first equation above and comparing gives us the constant K :

$$\left. \begin{aligned}x' &= -vt' \\x' &= 0 + K \cdot ct\end{aligned} \right\} K = -\left(\frac{v}{c}\right) \left(\frac{t'}{t}\right)\tag{3.2.2}$$

Events A and B both occur at the origin of the unprimed frame, so the time span between them is the proper time, and the frame is inertial, so it is the spacetime interval. Therefore the time measured between these events in the primed and unprimed frames are related according to the usual time dilation formula:

$$t' = \gamma_v t\tag{3.2.3}$$

Plugging this in above gives us the constant K :

$$K = -\left(\frac{v}{c}\right) \gamma_v\tag{3.2.4}$$

Using this same event B, we can obtain the constant M as well. Plugging in $x = 0$ gives:

$$ct' = 0 + M \cdot ct \Rightarrow M = \frac{t'}{t} = \gamma_v\tag{3.2.5}$$

Recapping what we have so far:

$$\begin{aligned}x' &= J \cdot x - \left(\frac{v}{c}\right) \gamma_v ct \\ct' &= L \cdot x + \gamma_v ct\end{aligned}\tag{3.2.6}$$

Now to determine the other two constants, define "event B" as occurring at the origin of the primed frame, $x' = 0$. The unprimed observer will see this event occur at the position $x = vt$, which we can plug back in to get:

$$0 = J \cdot vt - \gamma_v vt \Rightarrow J = \gamma_v \quad (3.2.7)$$

To find the final constant L requires noting that the time measured in the primed frame for event B is now the proper time, and a bit more algebra than was needed for the previous constants (which is omitted here):

$$\left. \begin{aligned} ct' &= L \cdot vt + \gamma_v ct \\ t &= \gamma_v t' \end{aligned} \right\} L = -\left(\frac{v}{c}\right) \gamma_v \quad (3.2.8)$$

Putting everything together gives us the **Lorentz transformation equations**:

$$\begin{aligned} ct' &= \gamma_v \left[ct - \left(\frac{v}{c}\right) x \right] \\ x' &= \gamma_v \left[x - \left(\frac{v}{c}\right) ct \right] \\ y' &= y \\ z' &= z \end{aligned} \quad (3.2.9)$$

The symmetry between the x and t variable is apparent, and shows the important difference between relativity and galilean physics – time is not universal and unaffected by the position of an event. Notice that when the velocity is very small compared to the speed of light (as it is in our everyday experience), then letting $\frac{v}{c} \rightarrow 0$ changes the Lorentz transformation equations into the Galilean transformation equations.

Finally, it should be noted that these transformations can also be written in terms of *changes* in these variables from one event to another. In effect, this is hidden in the equations themselves, as event A simply has all the variables equal to zero.

These equations give the spacetime coordinates of an event in the primed frame given the spacetime coordinates of the same event in the unprimed frame. But what if we want to do the reverse – find the coordinates of the event in the unprimed frame from those in the primed frame? [This is called the **inverse** of this transformation.] It's actually quite easy to do – the only difference in perspectives between these two frames is the sign of the velocity. We get the inverse transformation by simply replacing the v everywhere in the equations with $-v$.

Example 3.2.1

We have said that the interval-squared $\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$ is an invariant, which means that it is the same in every inertial frame. Use the Lorentz transformation equations to show that this is true.

Solution

We want to show that $\Delta s'^2 = \Delta s^2$, which makes this a pure plug-in. Clearly the y and z changes are equal in both frames, so we will ignore them and deal with just the t and x changes:

$$\begin{aligned} \Delta s'^2 &= c^2 \Delta t'^2 - \Delta x'^2 \\ &= \left(\gamma_v \left[c \Delta t - \left(\frac{v}{c}\right) \Delta x \right] \right)^2 - \left(\gamma_v \left[\Delta x - \left(\frac{v}{c}\right) c \Delta t \right] \right)^2 \\ &= \gamma_v^2 \left[\left(c^2 \Delta t^2 - 2v \Delta x \Delta t + \frac{v^2}{c^2} \Delta x^2 \right) - \left(\Delta x^2 - 2v \Delta x \Delta t + v^2 \Delta t^2 \right) \right] \\ &= \cancel{\gamma_v^2} \left[\left(1 - \frac{v^2}{c^2} \right) (c^2 \Delta t^2 - \Delta x^2) \right] \\ &= \Delta s^2 \end{aligned}$$

Revisiting Previous Results

After all that struggle with thought experiments and spacetime diagrams, only now do we have a simple, powerful tool for achieving the same results. Time dilation is downright trivial. If (unprimed) Ann sees two events occur at the same place ($\Delta x = 0$) separated by a time interval Δt , then the time span that (primed) Bob measures between these events is:

$$c \Delta t' = \gamma_v \left[c \Delta t - \left(\frac{v}{c}\right) \cancel{\Delta x} \right] \Rightarrow \Delta t' = \gamma_v \Delta t \quad (3.2.10)$$

We can also look at simultaneity. Events that are simultaneous in Ann's frame ($\Delta t = 0$) are not simultaneous in Bob's:

$$c\Delta t' = \gamma_v \left[c\cancel{\Delta t}^0 - \left(\frac{v}{c} \right) \Delta x \right] \neq 0 \quad (3.2.11)$$

Looking at this expression, we also see that $\Delta t'$ is negative (i.e. $t_2 < t_1$) when $\Delta x'$ is positive (i.e. $x_2 > x_1$). This means that for the two events that Ann sees as simultaneous, Bob sees the event with the greater x -value as occurring first (note that we are still assuming that Bob is moving in the $+x$ -direction relative to Ann). So is Ann flies by Bob in a spaceship where she sees lights on the front and rear of her ship flashing in sync, Bob sees the light on the rear of her ship flashing ahead of the light on the front.

Reproducing length contraction is a bit more difficult to obtain from the Lorentz transformation equations. the reason is that the length that is measured by one observer depends upon different events than the length measured by the other observer. That is, the length of an object in a given frame is the distance between events located at both ends of the object *that occur at the same time*, and as just noted, events simultaneous in one frame are not simultaneous in the other. Nevertheless, we can get the length contraction result with some care.

Two events that are simultaneous at both ends of an object according to Bob gives:

$$0 = c\Delta t' = \gamma_v \left[c\Delta t - \frac{v}{c} \Delta x \right] \Rightarrow c\Delta t = \frac{v}{c} \Delta x \quad (3.2.12)$$

Plugging this back into the transformation for the length measured by Bob gives the length contraction:

$$\Delta x' = \gamma_v [\Delta x - v\Delta t] = \gamma_v \left[\Delta x - v \left(\frac{v}{c} \Delta x \right) \right] = \gamma_v \Delta x \left[1 - \frac{v^2}{c^2} \right] = \frac{\Delta x}{\gamma_v} \quad (3.2.13)$$

This page titled [3.2: Lorentz Transformation](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [2.2: Lorentz Transformation](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

3.3: Velocity Addition

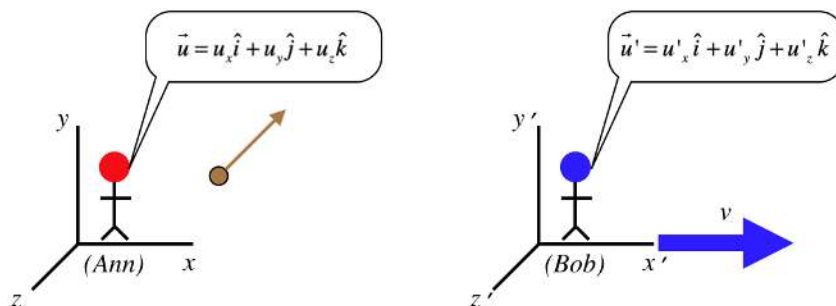
Cosmic Speed Limit

We have already said a few times that nothing we have found makes any sense if ordinary objects can move at or beyond the speed of light. So doesn't the whole theory come crashing down in the following scenario?

Bob is moving past Ann at $0.6c$ in the $+x$ -direction, as she fires off her powerful potato gun in the $-x$ -direction. Ann measures the speed of the potato in her frame to be $0.8c$. Bob now measures the speed of the potato – doesn't he see it moving at $1.4c$, ruining this whole crazy theory?

This sensible result comes straight from the equally-sensible (and yet wrong) Galilean transformation. Let's consider a case where both Ann and Bob are looking at the same object that is moving relative to both of them (as always, they are moving with a relative speed of v along the x -axis):

Figure 3.3.1 – Ann and Bob Watch the Same Ball



If we use the Galilean transformation to determine how these velocities $\vec{u}$ and $\vec{u}'$ relate to each other, we get:

$$\begin{aligned}
 t' = t & \Rightarrow \frac{dt'}{dt} = 1 \\
 x' = x - vt & \Rightarrow u'_x = \frac{dx'}{dt'} = \frac{dx'}{dt} \frac{dt}{dt'} = \left(\frac{dx}{dt} - v \right) (1) = u_x - v \\
 y' = y & \Rightarrow u'_y = \frac{dy'}{dt'} = \frac{dy'}{dt} \frac{dt}{dt'} = \left(\frac{dy}{dt} \right) (1) = u_y \\
 z' = z & \Rightarrow u'_z = \frac{dz'}{dt'} = \frac{dz'}{dt} \frac{dt}{dt'} = \left(\frac{dz}{dt} \right) (1) = u_z
 \end{aligned} \tag{3.3.1}$$

So this gives the result we expected before we ever heard of relativity. But now we know that the Galilean transformations are just an approximation of what really describes our universe, and that we actually need to use the Lorentz transformation equations instead. So let's follow precisely the method we followed above with the Lorentz transformation to find how velocities add in relativity. [Note that we have to use the γ that goes with the relative velocity of the two frames (since that is what transforms the coordinates), not the gamma related to the relative motion of the ball. This velocity is not changing, so γ_v is treated as a constant in the derivatives to come.]

$$\begin{aligned}
 ct' = \gamma_v \left[ct - \left(\frac{v}{c} \right) x \right] & \Rightarrow \frac{dt'}{dt} = \gamma_v \left[1 - \left(\frac{v}{c^2} \right) \frac{dx}{dt} \right] = \gamma_v \left(1 - \frac{u_x v}{c^2} \right) \\
 x' = \gamma_v (x - vt) & \Rightarrow u'_x = \frac{dx'}{dt'} = \frac{dx'}{dt} \frac{dt}{dt'} = \frac{u_x - v}{1 - \frac{u_x v}{c^2}} \\
 & = \gamma_v \left(\frac{dx}{dt} - v \right) \left[\gamma_v \left(1 - \frac{u_x v}{c^2} \right) \right]^{-1} \\
 y' = y & \Rightarrow u'_y = \frac{dy'}{dt'} = \frac{dy'}{dt} \frac{dt}{dt'} = \frac{u_y}{\gamma_v \left(1 - \frac{u_x v}{c^2} \right)} \\
 & = \left(\frac{dy}{dt} \right) \left[\gamma_v \left(1 - \frac{u_x v}{c^2} \right) \right]^{-1} \\
 z' = z & \Rightarrow u'_z = \frac{dz'}{dt'} = \frac{dz'}{dt} \frac{dt}{dt'} = \frac{u_z}{\gamma_v \left(1 - \frac{u_x v}{c^2} \right)} \\
 & = \left(\frac{dz}{dt} \right) \left[\gamma_v \left(1 - \frac{u_x v}{c^2} \right) \right]^{-1}
 \end{aligned} \tag{3.3.2}$$

Note that the *signs* of the components of $\vec{u}$ are important here. When $u_x > 0$, the x -component of the ball's velocity measured in Ann's frame is in the same direction as Bob is moving relative to Ann, and Bob sees the velocity of the ball as a little bit faster than the Galilean transformation predicts, because the denominator of the velocity transformation is less than 1. On the other hand, if the ball's velocity has an x -component in Ann's frame that is opposite to Bob's direction of motion ($u_x < 0$), then the velocities add, but the sum is not as great as is predicted by the Galilean transformation, because the denominator is greater than 1.

So let's try the example that started this discussion. Putting in the velocity of the potato in Ann's frame and Bob's relative velocity with Ann gives the speed measured by Bob:

$$u'_x = \frac{u_x - v}{1 - \frac{u_x v}{c^2}} = \frac{-0.8c - 0.6c}{1 - \frac{(-0.8c)(0.6c)}{c^2}} = \frac{-1.4c}{1 + 0.48} = -0.946c \quad (3.3.3)$$

The cosmic speed limit is obeyed!

Example 3.3.1

Here's another idea to bring relativity to its knees: Maybe we can't get a speed greater than c for a moving **object** by stacking one velocity on top of another because neither of them can be going a speed c to begin with, but what if Ann shines a light so that the light's speed adds to Bob's speed? Won't Bob measure the light going faster than c ?

Solution

Well, this clearly violates the postulate of relativity, but let's use this idea to check our velocity formula anyway. Putting $u_x = \pm c$ into the velocity transformation indeed gives us the right answer:

$$u'_x = \frac{\pm c - v}{1 - \frac{(\pm c)(v)}{c^2}} = \pm c$$

Velocity Vectors

We have focused on the effects of velocity addition along the direction of motion, but perhaps a result we didn't see coming was that relative motion along the x -direction also affects the relationship between the two observers' measurements of velocities in directions perpendicular to the relative motion. This comes about because of the time dilation effect – if time is passing more slowly in another frame than in your own, then things are moving more slowly in those perpendicular directions. For example, if Bob sees time passing more slowly for Ann as she moves by, then when she drops her pencil, it will take longer to get to the floor from Bob's perspective than from Ann's perspective, so for Bob the pencil is moving more slowly, even though it is moving in a direction perpendicular to the relative motion.

Suppose as Bob zooms by, Ann fires a laser in the $+y$ direction. What does Bob see for the laser's speed and direction? Well, he had better see a speed of c , and thanks to the velocity transformation, he does:

$$\left. \begin{aligned} u_x = 0 &\Rightarrow u'_x = \frac{0 - v}{1 - 0} = -v \\ u_y = c &\Rightarrow u'_y = \frac{c}{\gamma_v (1 - 0)} = \sqrt{c^2 - v^2} \end{aligned} \right\} |\vec{u}'| = \sqrt{u'^2_x + u'^2_y} = c \quad (3.3.4)$$

Bob and Ann don't agree on the *direction* however. While Ann sees the laser going in the $+y$ direction, Bob sees an angle with his y' -axis:

$$\sin \theta = \frac{v}{c} \Rightarrow \theta = \sin^{-1} \left(\frac{v}{c} \right) \quad (3.3.5)$$

This page titled 3.3: Velocity Addition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Tom Weideman directly on the LibreTexts platform.

- 2.3: Velocity Addition by Tom Weideman is licensed CC BY-SA 4.0. Original source: [native](#).

3.4: Electricity and Magnetism

Motion in E&M

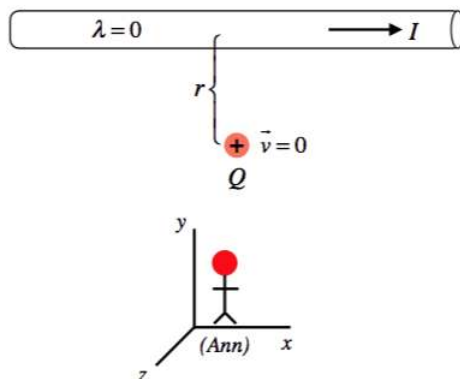
Electricity and magnetism was the main wildly-successful theory at the time that Einstein came onto the scene. Maxwell had shown how the two different forces could be unified into one mathematical theory. Electric charges cause both electric and magnetic fields, both of the fields exert forces on electric charges (in different ways), *and* each of the two fields (when they vary in time) can induce the existence of the other.

But when Einstein pondered his ideas about relative motion, an important question came to mind. The velocity of an electric charge appears in two different places in Maxwell's theory: A moving charge produces a magnetic field that is proportional to its speed, and a moving charge experiences a force from an external magnetic field that is proportional to its speed. So the obvious question to ask is, "What happens if we just view the moving charge from a frame that is stationary relative to the charge?" In this frame, it would seem that there is no magnetic field produced by the charge, and this charge would experience no force from an external magnetic field. Reconciling this is vital to confirming the validity of both theories, and we will do so with – what else? – a thought experiment.

Lorentz Force in Two Frames

Let us consider the following simple example from Physics 9C:

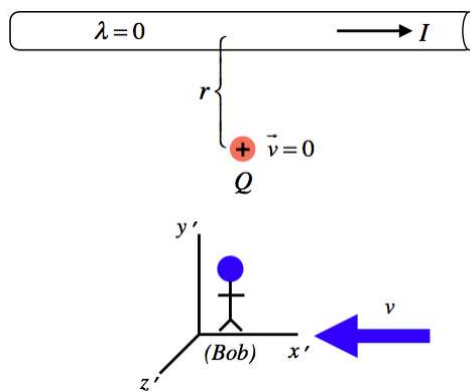
Figure 3.4.1 – Ann Observes a Stationary Charge Near a Current-Carrying Wire



Ann observes a positive point charge Q that is at rest in her inertial frame. Close by (a distance r away), parallel to the x -axis, is a long-straight wire that carries a current I in the $+x$ -direction. Like any other current-carrying wire, this one has zero electric charge density. It should be immediately clear that Ann will witness no net force on this charge – there is no charge density on the wire to produce an electric field, and although the current produces a magnetic field at the position of the charge, the charge isn't moving, so there is no magnetic force on it either.

Okay, all simple stuff. We continue with our thought experiment as we usually do, by considering what is seen by a different observer, Bob. Recall that electric current is typically due to movement of electrons in a conductor (while the protons remain stationary), and these electrons, being negatively-charged move in the direction opposite to the electric current. So for our thought experiment, let's choose Bob to be an observer in a frame that moves along with the electrons in the conductor:

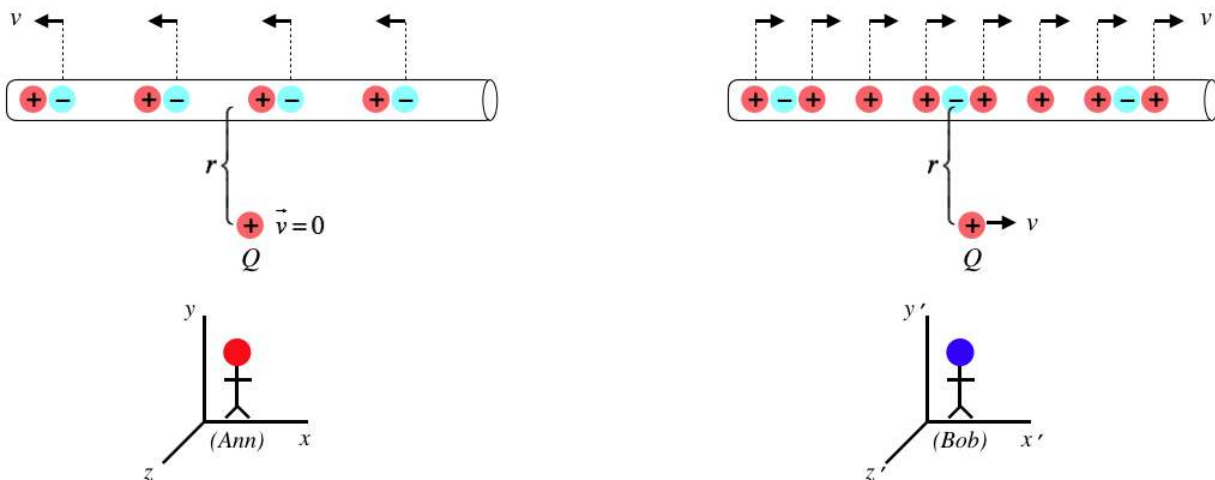
Figure 3.4.2 – Bob Moves by the Charge with the Electrons



At first blush, it might seem that Bob should see a magnetic force on the charge, because according to him, the electric charge is moving in the magnetic field of the wire. Specifically, the right-hand-rule tells us that the magnetic field of the current points into the page at the position of the charge, and the charge (according to Bob) is moving to the right. So applying the RHR for the Lorentz force $\vec{F} = Q\vec{v} \times \vec{B}$, we find that the force on this positive charge must be toward the wire. But if this force exists, then Bob would see the charge start moving closer to the wire, while Ann would see the charge remain stationary. Both of these results cannot be true, so what are we missing?

To answer this, we need to have a closer look at the charges that comprise the electric current. We have set things up so that Ann is stationary relative to the protons (with the electrons moving left at v), and Bob is stationary relative to the electrons, which means that relative to him, the protons are moving to the right at a speed v . At the start of our thought experiment, we said that Ann witnesses zero charge density on the wire. This is because the electrons provide a negative charge density that is equal in magnitude to the positive charge density provided by the protons. These densities are determined by the spacings of the charges – the closer together they are, the more dense the charge distribution is. But relative motions changes distances with length contraction, so Bob will not measure the charge separations to be the same that Ann measures.

Figure 3.4.3 – Comparing Charge Densities Measured



Here we see side-by-side the views of each observer in their rest frames. The electrons are moving relative to Ann, so to her their separation is length-contracted. In Bob's frame their separations are not length-contracted, so he sees them to be farther apart. Similarly, the protons are moving in Bob's frame, and stationary in Ann's so Bob sees them to be closer together than Ann. The end result is that while Ann sees zero charge density on the wire, Bob does not, because he sees the protons much more closely-spaced than the electrons. Interestingly, *both* observers see an electric current going to the right (positive charges to the right and negative charges moving to the left both result in current to the right).

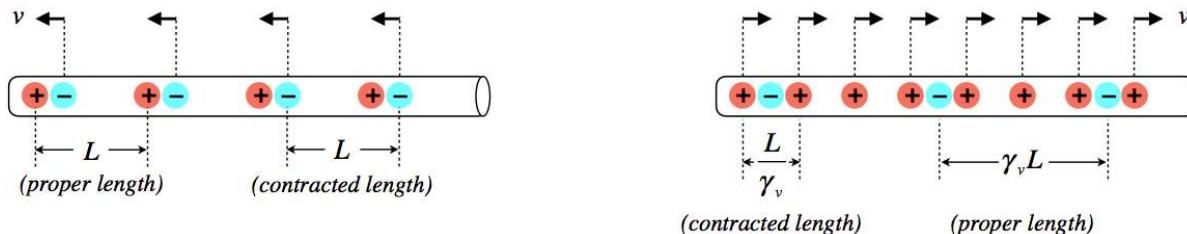
Without yet doing any math, we can begin to see why Bob's conclusion could possibly agree with Ann's. With a current to the right and a right-moving point charge, as we noted above, Bob does witness a magnetic force on the point charge toward the wire. But now we see that there is also an electric force due to the net positive charge density on the wire. This will of course repel the

positive point charge, so if the magnitude of this repulsive electric force happens to equal the magnitude of the attractive magnetic force, then Bob and Ann will agree, as they must!

Now for the Math

In *principle* the electric and magnetic forces on the point charge witnessed by Bob can cancel, but for us to actually confirm that special relativity is consistent with E&M, the math has to work out. There are a number of parts to put together here. We'll start with the charge density. We'll call the separation of neighboring charges in Ann's frame L , and adjust the separations in Bob's frame by length contraction:

Figure 3.4.4 – Charge Densities in Both Frames



The linear density is the charge per unit length, so:

$$\lambda_{Ann} = \frac{+q}{L} + \frac{-q}{L} = 0 \quad \lambda_{Bob} = \frac{+q}{\left(\frac{L}{\gamma_v}\right)} + \frac{-q}{\gamma_v L} = \gamma_v \frac{q}{L} \left(1 - \frac{1}{\gamma_v^2}\right) = \gamma_v \frac{q}{L} \left(\frac{v}{c}\right)^2 \quad (3.4.1)$$

From Physics 9C, we have a [formula](#) for the electric field of a line of charge, in terms of its line density. The electric field seen by Ann is clearly zero, and for Bob it is:

$$E_{Bob} = \frac{\lambda}{2\pi\epsilon_0 r} = \left(\frac{v^2}{c^2\epsilon_0}\right) \left(\frac{\gamma_v q}{2\pi r L}\right) \quad (3.4.2)$$

Okay, now we need the magnetic field. We start by reminding ourselves how to get the current from a moving charge density:

$$I = \frac{dq}{dt} = \frac{dq}{dx} \frac{dx}{dt} = \lambda v \quad (3.4.3)$$

Note that for Bob, only the positive charges are moving, so the λ that appears in this formula is the charge density of the positive charges only, not the total charge density. So the current is, according to Bob:

$$I = \left(\gamma_v \frac{q}{L}\right) v \quad (3.4.4)$$

We also have a [formula](#) for the magnetic field from the current of a long straight wire from Physics 9C, and plugging this current into it gives the magnetic field seen by Bob:

$$B_{Bob} = \frac{\mu_0 I}{2\pi r} = (\mu_0 v) \left(\frac{\gamma_v q}{2\pi r L}\right) \quad (3.4.5)$$

All that remains is to compute the Lorentz force measured by Bob. We have already establish that the electric and magnetic forces are in opposite directions, and since the motion of the point charge is perpendicular to the magnetic field, we have:

$$\begin{aligned} \vec{F} &= Q\vec{E} + Q\vec{v} \times \vec{B} \Rightarrow F_{Bob} = Q \left(\frac{v^2}{c^2\epsilon_0}\right) \left(\frac{\gamma_v q}{2\pi r L}\right) - Qv (\mu_0 v) \left(\frac{\gamma_v q}{2\pi r L}\right) \\ &= \left(\frac{1}{c^2\epsilon_0} - \mu_0\right) \left(\frac{\gamma_v v^2 q Q}{2\pi r L}\right) \end{aligned} \quad (3.4.6)$$

And now the calculation is completed at last using another [result](#) from Physics 9C:

$$c^2 = \frac{1}{\mu_0\epsilon_0} \Rightarrow F_{Bob} = 0 \quad (3.4.7)$$

Full Unification Of Electricity and Magnetism

What this thought experiment shows us is the remarkable unification of the electric and magnetic forces started by Maxwell and completed by Einstein. Maxwell showed that a single source (electric charge) is responsible for both fields and is affected by both fields. With special relativity, Einstein went a step further, and showed that both fields are actually a single field, which has electric and magnetic properties that will be manifested differently depending upon the inertial frame in which they are observed.

It is actually possible to derive specifically how electric and magnetic fields mix into each other as one changes reference frames, in the same way that the Lorentz transformations describe how position and time variables mix into each other. We won't reproduce the derivation here, but here is the final result:

$$\begin{aligned}\vec{E}'_{\parallel} &= \vec{E}_{\parallel} & \vec{B}'_{\parallel} &= \vec{B}_{\parallel} \\ \vec{E}'_{\perp} &= \gamma_v \left(\vec{E}_{\perp} + \vec{v} \times \vec{B} \right) & \vec{B}'_{\perp} &= \gamma_v \left(\vec{B}_{\perp} - \frac{1}{c^2} \vec{v} \times \vec{E} \right)\end{aligned}\tag{3.4.8}$$

In these formulas, the quantities require some explanation. The primed frame is moving with a velocity $\vec{v}$ relative to the unprimed frame. The field vectors with no subscripts are the full field vectors, while those with the " $\parallel$ " subscript are the field components parallel to $\vec{v}$, and those with the " $\perp$ " subscript are the field components perpendicular to $\vec{v}$.

This page titled [3.4: Electricity and Magnetism](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

- [2.3: Velocity Addition](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

CHAPTER OVERVIEW

4: Dynamics in Special Relativity

[4.1: Momentum Conservation](#)

[4.2: Energy Conservation](#)

This page titled [4: Dynamics in Special Relativity](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

4.1: Momentum Conservation

Another Thought Experiment

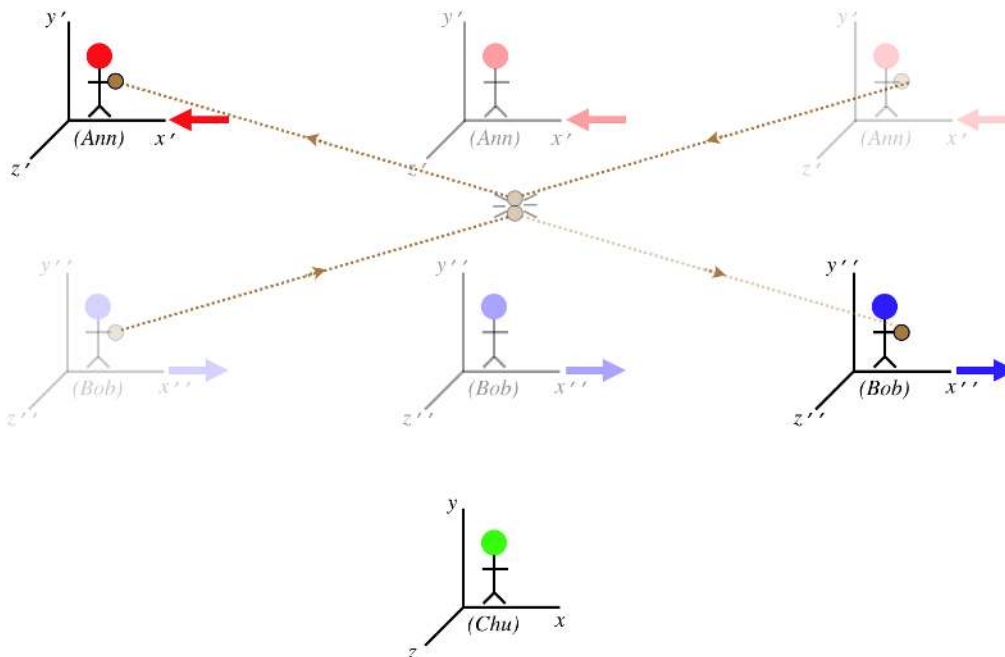
We have completed our exploration of the fundamentals of relativistic kinematics and its consequences. Now it is time to examine the consequences of the relativity principle in the area of dynamics. Our first clue that something needs to be done differently comes from the basis of all dynamics: Newton's laws of motion. In particular, the second law looks like:

$$\vec{F}_{net} = \frac{d\vec{p}_{cm}}{dt} \quad (4.1.1)$$

The obvious question is, with respect to what time measurement is this momentum changing? This is actually a very difficult problem to deal with in relativity in that it is difficult to apply the relativity principle, so we will instead look at a consequence of this law – momentum conservation. If someone in one frame observes a collision of two objects and declares that momentum is conserved, then an observer in another frame watching the same collision should conclude the same thing. Momentum conservation is among the most cherished principles of physics, and if an experiment could be performed where two inertial observers do not agree that it is upheld, then that would cause problems for the relativity principle. This calls for a thought experiment!

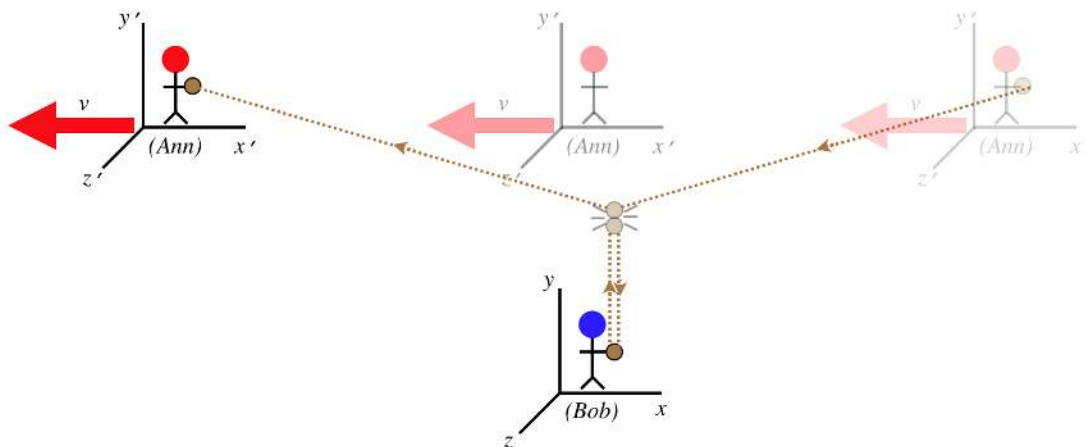
We have three players involved here (Ann, Bob, and Chu), and we are initially viewing the activities from Chu's perspective, who sees Ann moving in the $-x$ -direction and Bob moving in the $+x$ -direction at equal speeds. Ann and Bob both throw identical balls in a direction that is (from their own perspectives) along their y axes, at speeds they each measure to be u . The balls collide elastically with each other, and return to where they started.

Figure 4.1.1 – Chu Observes Collision



From Chu's perspective, everything is completely symmetric, and so he concludes that the momentum (and most notably, the y -component of the momentum) in this collision is conserved. Okay, let's confirm that all observers agree upon this basic physics principle by looking at the very same collision from Bob's perspective. To Bob, his ball's motion parallels his y -axis, while Ann's has an x -component (equal to her frame's relative motion).

Figure 4.1.2 – Bob Observes Collision



Bob now compares the y -components of velocity of the two balls. We already stated that Bob measures the velocity of his ball to be u , and Ann measures velocity of her ball to be the same, but because of the strangeness of velocity, neither agrees that the *other* ball is moving that fast. Bob says that the y -component of velocity of Ann's ball is (u_x and u_y are the x and y components of Ann's ball, respectively, according to Ann):

$$u'_y(\text{Ann's ball}) = \frac{u_y}{\gamma_v(1 - \frac{u_x v}{c^2})} = \frac{u}{\gamma_v(1 - 0)} = \frac{u}{\gamma_v} \quad (4.1.2)$$

This is *smaller* than the vertical speed that Bob measures for his own ball, which means that Ann's ball (which has the same mass as his ball) had to have less momentum going into the collision than his ball. But he sees his own ball come back at the same speed that it left, and the collision is elastic (his ball isn't any warmer), and these are properties of a head-on collision with an identical ball moving at the same speed in the opposite direction if momentum is conserved. So something is wrong here.

A New-and-Improved Momentum

There are only three possible conclusions we can draw from the discrepancy shown in this thought experiment:

- Momentum conservation is not a fundamental principle of physics, since one observer measuring no change in momentum (in the example above, Chu) for a closed system does not ensure that every other observer (Ann and Bob) gets that same result.
- Mass is measured differently in different reference frames. If Bob measures the mass of Ann's ball to be greater than his own, then that could compensate for the lower velocity so that their two balls once again have equal y -components of momenta.
- Our definition of momentum, while useful for low velocities, needs a facelift to handle relativistic speeds.

For physicists, stopping at the first of these was never an option – conservation principles and momentum in particular had been revered for far too long. In the early years (and for quite some time afterward), the second option was the accepted explanation, and it works fine. In modern times, however, the physics community has instead embraced the third option – we prefer to characterize mass as an invariant quantity that is inherent to matter, and simply admit that our original definition of momentum was insufficient.

Deriving the correct form of the momentum is challenging until we get to 4-vectors in a future chapter, so right now we will just be given this formula and confirm that it works for the Ann/Bob/Chu thought experiment above. For an object moving at a speed u relative to an observer, the momentum of the object in the frame of that observer is defined to be:

$$\vec{p} \equiv m\gamma_u \vec{u} \quad (4.1.3)$$

Alert

Up to now, we have usually dealt with γ_v that relates the frames of two observers, whose relative speed we define as ' v ,' but here the γ_u relates the frame of the observer to that of the moving object, not another observer. As we will be relating the momenta in frames of observers, it will be important to keep straight the difference between γ_v and γ_u . As one example, if we have two observers, the γ_v that relates their frames are the same for both of them, but the γ_u that one observer uses for a moving object is not the same as the γ_u used by the other observer for the same moving object, since $u \neq u'$.

As a first check, it is clear that at slow velocities, momentum reduces to our usual definition of momentum, since for $u \ll c$, $\gamma_u \rightarrow 1$.

Proving that this form of momentum is conserved in the thought experiment above requires careful accounting of velocities, as there are many involved here. In an effort to not have to carry primes on all our variables through the calculation, we will look at this collision from Ann's perspective – clearly the collision has all the same features for her that it has for Bob. We will use the subscript 'A' whenever referring to a quantity specifically related to Ann's ball, and use a 'B' for Bob's ball. We seek to write all of these quantities in terms of the value u , the speed that each measures for their own ball; and v , the relative speed of their two frames.

We need the γ 's for the two balls from Ann's perspective in order to compare momenta in her frame. The γ for her ball is easy, as she sees it moving with a speed of u :

$$\vec{p}_A = -m \gamma_{uA} u_A \hat{j} = -\frac{m u}{\sqrt{1 - \frac{u^2}{c^2}}} \hat{j} \quad (4.1.4)$$

When Ann looks at the ball thrown by Bob, she sees that it has an x -component equal to $+v$, and its y -component is determined from the velocity addition formula. Bob sees no x' component for his ball's velocity $u'_{Bx} = 0$, and he measures the y' component for his ball's velocity to be $u'_{By} = u$, so using the velocity addition formula to determine the two components of velocity of Bob's ball in Ann's frame gives:

$$\begin{aligned} u_{Bx} &= \frac{u'_{Bx} + v}{1 + \frac{u'_{Bx} v}{c^2}} = \frac{0 + v}{1 + 0} = v \\ u_{By} &= \frac{u'_{By}}{\gamma_v \left(1 + \frac{u'_{Bx} v}{c^2}\right)} = \frac{u}{\gamma_v (1 + 0)} = \frac{u}{\gamma_v} \end{aligned} \quad (4.1.5)$$

We now need to construct the γ_u for Bob's ball as seen by Ann, which means we first need to construct the square of the speed for Bob's ball according to Ann:

$$u_B^2 = u_{Bx}^2 + u_{By}^2 = v^2 + \frac{u^2}{\gamma_v^2} \Rightarrow \gamma_{uB} = \frac{1}{\sqrt{1 - \frac{u_B^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2} - \frac{u^2}{\gamma_v^2 c^2}}} \quad (4.1.6)$$

Plugging this in gives the y -component of momentum of Bob's ball as measured by Ann, which we can then compare to the momentum of Ann's ball (which is entirely in the y -direction):

$$p_{By} = \gamma_{uB} m u_{By} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2} - \frac{u^2}{\gamma_v^2 c^2}}} m \frac{u}{\gamma_v} \quad (4.1.7)$$

Showing that momentum is conserved with its new definition is now a matter of showing that this equals the magnitude of Equation 4.1.4:

$$\frac{1}{\sqrt{1 - \frac{v^2}{c^2} - \frac{u^2}{\gamma_v^2 c^2}}} \cancel{m} \cancel{u} = \frac{\cancel{m} \cancel{u}}{\sqrt{1 - \frac{u^2}{c^2}}} \quad (4.1.8)$$

Inverting and squaring both sides of the equation and simplifying completes the proof:

$$\left(1 - \frac{v^2}{c^2} - \frac{u^2}{\gamma_v^2 c^2}\right) \gamma_v^2 = 1 - \frac{u^2}{c^2} \Rightarrow \left(1 - \frac{v^2}{c^2}\right) \gamma_v^2 - \frac{u^2}{c^2} = 1 - \frac{u^2}{c^2} \quad (4.1.9)$$

Due to the symmetry of the two cases, it should be clear that Bob will get the same result, and from Chu's perspective, the new version of momentum changes the magnitude of the momenta of both balls equally, so he will naturally again witness momentum conservation. Indeed, with this new definition of momentum, every frame will agree that it is conserved before and after the collision.

This page titled [4.1: Momentum Conservation](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [2.4: Momentum Conservation](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

4.2: Energy Conservation

One Last Thought Experiment

With an understanding of relativistic momentum conservation now firmly in hand, we will have a look at an inelastic collision. When we first encountered these in Physics 9A, we said that no energy was actually "lost," it was just converted into another form (thermal). Given that thermal energy is at its core "microscopically mechanical," this assessment of it changing form is really just a shortcut, and in fact the only things that change in terms of energy is how it is packaged. For example, we could model a simple (perfectly) inelastic collision between two equal masses this way:

Figure 4.2.1 – A Simple Model of an Inelastic Collision



In this collision, if we can see what is going-on inside the boxed system after the collision, we can account for all of the incoming energy – part of it goes to the kinetic energy of the boxed system, and part of it to the potential and kinetic energy associated with the oscillation of the two particles. If we can't see what's going on, then we can only see the kinetic energy of the boxed system, and we call the leftover energy "internal energy" within the boxed system.

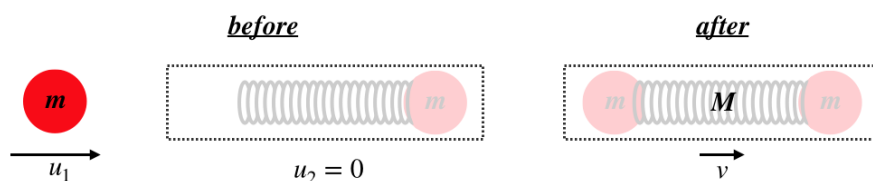
In Physics 9A, for the non-relativistic case of a collision of this kind, we showed that the fraction of the initial kinetic energy the box system has after the collision is given by:

$$\frac{KE_{after}}{KE_{before}} = \frac{m_1}{m_1 + m_2} = \frac{1}{2} \quad (4.2.1)$$

Put another way, for this case (non-relativistically, when the masses are equal), the energy contained in the oscillation of the two masses equals the kinetic energy of the box system after the collision.

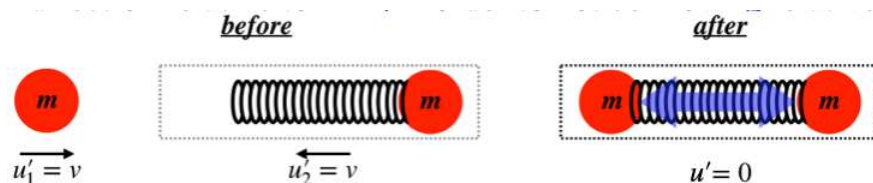
Let's see how all this works out for relativity with our new definition of momentum. We will watch this collision from two different perspectives. The first is Ann, who sees the collision from the perspective shown above, and who cannot see inside the boxed system, so she doesn't even know the system's mass after the collision (she calls it ' M '). Here is the before/after diagram she uses for momentum conservation:

Figure 4.2.2 – Ann's Before/After Diagram for a Perfectly Inelastic Collision



Bob will view the very same collision from another frame that is moving to the right with a speed of v , which is in the rest frame of the system after the collision. Unlike Ann, we'll say that he is able to see what is going on inside the box. With the two parts having equal mass, and coming to rest after the collision, he naturally must see both halves moving at the same speed, so from his perspective, the collision looks like this:

Figure 4.3.3 – Bob's Before/After Diagram for a Perfectly Inelastic Collision



Okay, so let's invoke relativistic momentum conservation for Ann. If she uses this to determine the unknown mass M , she finds:

$$\gamma_{u_1} m u_1 + 0 = \gamma_v M v \Rightarrow M = \left(\frac{\gamma_{u_1}}{\gamma_v} \right) \left(\frac{u_1}{v} \right) m \quad (4.2.2)$$

Next all we have to do is relate u_1 and v to each other using velocity addition:

$$u_1 = \frac{u'_1 + v}{1 + \frac{u'_1 v}{c^2}} = \frac{2v}{1 + \frac{v^2}{c^2}} \Rightarrow \frac{u_1}{v} = \frac{2c^2}{c^2 + v^2} \quad (4.2.3)$$

Writing γ_{u_1} in terms of v :

$$\begin{aligned} \gamma_{u_1} &= \frac{1}{\sqrt{1 - \frac{u_1^2}{c^2}}} = \frac{1}{\sqrt{1 - \left(\frac{u_1}{v} \cdot \frac{v}{c} \right)^2}} = \frac{1}{\sqrt{1 - \left(\frac{2c^2}{c^2 + v^2} \cdot \frac{v}{c} \right)^2}} \\ &= \frac{c^2 + v^2}{\sqrt{c^4 - 2v^2c^2 + v^4}} = \frac{c^2 + v^2}{c^2 - v^2} \end{aligned} \quad (4.2.4)$$

Plugging everything back into Equation 4.2.2 reveals the mass that Ann measures for the combined system:

$$M = \left(\frac{\frac{c^2 + v^2}{c^2 - v^2}}{\gamma_v} \right) \left(\frac{2c^2}{c^2 + v^2} \right) m = 2\gamma_v m \quad (4.2.5)$$

Wait, Ann finds that the mass of the combined system is actually *larger* than $2m$? This doesn't seem to agree with what Bob knows about the combined system. If the two masses were just held together, then Bob would be certain that the combined system would have a mass of $2m$, but perhaps there is something about the fact that the masses are oscillating on the spring that adds to the mass of the system?

Perhaps we can get a clue about what is going on by looking at the mass discrepancy in the case of our familiar slow-moving world. The apparent additional mass is:

$$\Delta m = M - 2m = M - \frac{M}{\gamma_v} = M \left(1 - \frac{1}{\gamma_v} \right) = M \left(1 - \sqrt{1 - \frac{v^2}{c^2}} \right) \quad (4.2.6)$$

Now using the usual $v \ll c$ approximation:

$$\begin{aligned} \sqrt{1 - \delta} &\approx 1 - \frac{1}{2}\delta \Rightarrow 1 - \sqrt{1 - \frac{v^2}{c^2}} \approx 1 - \left(1 - \frac{1}{2} \frac{v^2}{c^2} \right) \\ &= \frac{1}{2} \frac{v^2}{c^2} \Rightarrow \Delta m \approx \frac{1}{2} M \frac{v^2}{c^2} \end{aligned} \quad (4.2.7)$$

For this low-velocity case of two equal masses, we said above that the kinetic energy of the box system after the collision equals the internal energy contained in the oscillations, we therefore have:

$$\text{internal energy} = \frac{1}{2} M v^2 = \Delta m c^2 \quad (4.2.8)$$

It appears that the energy that starts as kinetic and becomes internal due to an inelastic collision is manifested – according to relativity – as an increase in the mass of the system where the internal energy is contained, with a conversion factor of c^2 .

Total, Kinetic, and Rest Energy

What constitutes "internal energy" is determined by what we define as a system: Just aggregate a group of particles, and that group's collective mass is not simply the sum of the masses of the particles in the group, but must also include the mass that is equivalent to the energy of all their internal motion and interactions according to $E = mc^2$. This famous equation is known as *mass-energy equivalence*, and it has interesting implications. For example, if we make an object hotter, then it contains more internal energy and therefore has more mass than when it is cooler.

In the example above, let's suppose Bob can't see inside the box. As we have said before, observers agree on masses, so he and Ann agree that the box has a mass of M . The box isn't moving in Bob's frame, so internal energy is the *only* energy the box has. The energy of a system measured in its rest frame is called the *rest energy*, and it comes from the system's total mass and mass-energy equivalence:

$$E_{\text{rest}} = M c^2 \quad (4.2.9)$$

If the energy of a system is instead measured in a frame in which it is not at rest, then there is a kinetic energy component that needs to be added to the rest energy to get the total energy. Naturally the total energy of a given system will be greater in frames in which the velocity of the system is greater. We can write the total energy as an unknown function of the velocity of the system in the frame, multiplied by the rest energy:

$$E_{tot} = f(u) mc^2 \quad (4.2.10)$$

To obtain this function, let's look at the collision above from Bob's perspective. The energy is conserved, and at the end it is just the rest energy. Before the collision, the two particles have equal total energies whose sum is the final energy:

$$f(v) mc^2 + f(v) mc^2 = Mc^2 \Rightarrow f(v) = \frac{M}{2m} \Rightarrow f(v) = \gamma_v \quad (4.2.11)$$

The final equality comes courtesy of [Equation 4.2.5](#). So we conclude that the total energy of an object with mass m moving at a speed of u is given by:

$$E_{tot} = \gamma_u mc^2 \quad (4.2.12)$$

Example 4.2.1

Show that energy is conserved for the collision above when measured in Ann's frame.

Solution

The total energy of the system before the collision comes in two pieces – the total energy of the incoming mass, and the rest energy of the target mass. After the collision the system's energy consists of the total energy of the moving total mass. We seek to show that these are equal:

$$\gamma_{u_1} mc^2 + mc^2 \stackrel{?}{=} \gamma_v Mc^2 \Rightarrow \gamma_{u_1} + 1 \stackrel{?}{=} \gamma_v \frac{M}{m}$$

Now substitute for $\frac{M}{m}$ using [Equation 4.2.5](#), giving:

$$\gamma_{u_1} + 1 \stackrel{?}{=} 2\gamma_v^2$$

Now use [Equation 4.2.4](#) to put everything in terms of v and c , and do the algebra:

$$\begin{aligned} \frac{c^2 + v^2}{c^2 - v^2} + 1 &= \frac{2}{1 - \frac{v^2}{c^2}} \\ \frac{c^2 + v^2 + c^2 - v^2}{c^2 - v^2} &= \frac{2c^2}{c^2 - v^2} \end{aligned}$$

With the total energy and the rest energy of a system now in hand, it is easy to define the kinetic energy as the difference of the two:

$$KE = (\gamma_u - 1) mc^2 \quad (4.2.13)$$

Example 4.2.2

Show that the relativistic kinetic energy is consistent with the non-relativistic definition of kinetic energy for speeds much less than c .

Solution

Whenever we see the phrase "speeds much less than c ," we immediately think of expanding γ to first order in $\frac{v^2}{c^2}$, as we did in [Equation 2.5.7](#):

$$KE = [\gamma_u - 1] mc^2 = \left[\left(1 - \frac{u^2}{c^2} \right)^{-\frac{1}{2}} - 1 \right] mc^2 \approx \left[\left(1 + \frac{1}{2} \frac{u^2}{c^2} \right) - 1 \right] mc^2 = \frac{1}{2} mu^2$$

Combining Energy and Momentum

Back in Physics 9A, we found a very useful formula that relates kinetic energy to momentum. It's clear that the same formula does not work for relativity:

$$KE = (\gamma_u - 1) mc^2$$

$$\frac{p^2}{2m} = \frac{1}{2m} (\gamma_u mu)^2 = \frac{1}{2} \gamma_u^2 mu^2 \quad (4.2.14)$$

This doesn't mean that there is no formula that relates these two quantities. Indeed:

$$E^2 = \gamma_u^2 m^2 c^4 = \frac{c^2}{c^2 - u^2} m^2 c^4 = \left(\frac{u^2}{c^2 - u^2} + 1 \right) m^2 c^4 = \frac{c^2}{c^2 - u^2} m^2 u^2 c^2 + m^2 c^4$$

$$= \gamma_u^2 m^2 u^2 c^2 + m^2 c^4 = p^2 c^2 + m^2 c^4 \quad (4.2.15)$$

So the alternative ways of writing the total energy are:

$$E = \gamma_u mc^2 = \sqrt{p^2 c^2 + m^2 c^4} \quad (4.2.16)$$

Massless Particles

With the γ_u multiplying mc^2 in the energy equation, we have another reason to insist that the speed of light is unobtainable – for a system to attain the speed of light, it would need to acquire infinite energy. But if this is true, does that mean that light has infinite energy? Of course not – we can measure the energy in light by absorbing it in matter and measuring the temperature change of the matter. So then how does light get away with moving at the cosmic speed limit? The answer is that while γ_u for light goes to infinity, the mass of a light "particle" (called a *photon*) turns out to be zero. The product of these two numbers turns out to result in a finite value.

Using the other energy equation tells us even more. Setting the mass equal to zero gives us a very simple relationship between the energy of a photon and its momentum:

$$E_{\text{photon}} = pc \quad (4.2.17)$$

So yes, light has both energy *and* momentum. Again, it might seem strange that something without mass can have momentum, but with γ_u exploding to infinity and the mass vanishing, this is again possible. The difference between light and matter in this regard is that photons don't have any rest energy – all of the energy comes from its momentum.

This page titled [4.2: Energy Conservation](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#) directly on the LibreTexts platform.

- [2.5: Energy Conservation](#) by [Tom Weideman](#) is licensed [CC BY-SA 4.0](#). Original source: [native](#).

CHAPTER OVERVIEW

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

[5.1: Blackbody Radiation](#)

[5.2: The Photoelectric Effect](#)

[5.3: Compton Effect](#)

[5.4: Double-Slit Experiment](#)

[5.5: Matter Has Wave Properties, Too!](#)

[5.6: Summarizing this Wave/Particle Mess](#)

This page titled [5: "Wait, what?!" Experiments Reveal Cracks in Our Understanding](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

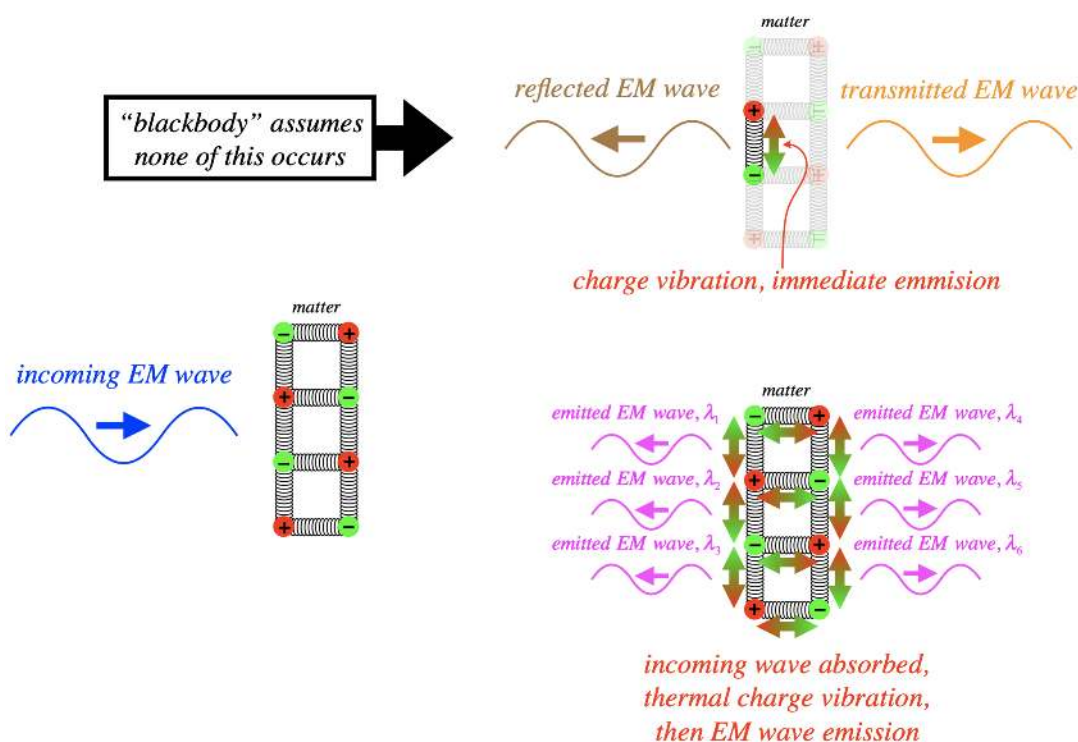
5.1: Blackbody Radiation

What is a "Blackbody"?

In Physics 9B we briefly discussed the idea of a *blackbody* when we discussed heat transfer by radiation. Now that you have taken Physics 9C, you can understand a little more detail about this concept, and see why it was causing physicists fits right at the turn of the century.

We know that EM radiation consists of fluctuating electric and magnetic fields. These fluctuations need to be at least of second-order (since the derivatives in the wave equation are second order), which means that they can only be created by accelerating electric charges. We often model electric charges in matter as being held in a lattice with spring-like bonds. So if those charges are oscillating in those bonds, they can give off light. There are essentially two ways that a piece of matter does this. One is for EM radiation to strike the matter and the fluctuating fields of the radiation cause the charges to vibrate, which in turn gives off more radiation. Another is to simply heat the matter, which we know causes the particles comprising it to vibrate. The former version is essentially just transmission or reflection. In the case of a blackbody, we ignore this transmission/reflection mode. Instead, we assume that all of the incoming EM radiation is totally absorbed by the matter, which then redistributes the energy throughout its mass in the form of thermal energy, and then the only EM radiation it emits is due to thermal vibrations of the charges.

Figure 5.1.1 – Blackbodies "Thermalize" Radiation

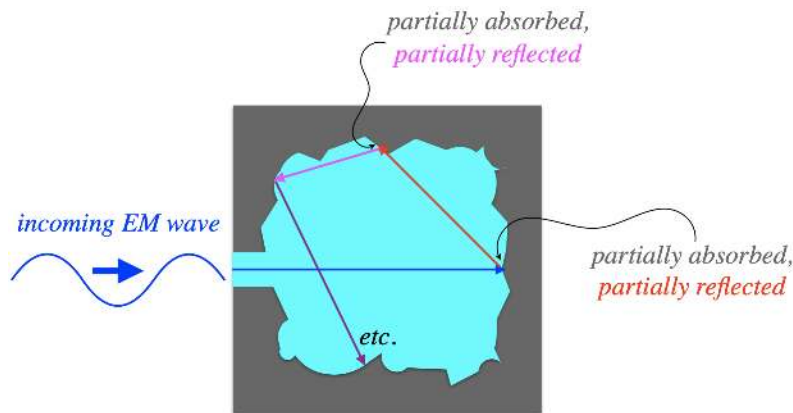


One might ask where the name "blackbody" comes from. Imagine some yellow light striking such an object. If it is totally reflected, then yellow light is emitted by it. But if the light's energy is absorbed by the object, and that energy is distributed randomly through the atoms which then emit light in a wide spectrum of frequencies, then the light that comes out will not all be in the visible spectrum, and if the amount in the visible spectrum is small, then the object's appearance is black.

Of course, an object's ability to turn incoming light into randomly-partitioned energy amongst its atoms depends upon a number of factors. In Physics 9B, we lumped all those factors into a *single factor we called the emissivity*. Mostly this quantity depends upon the surface properties of the object, as this is where reflection occurs, and reflection needs to be avoided for the incoming light's energy to be absorbed and distributed within the object. As we stated in Physics 9B, a blackbody is an object with emissivity of 1, which means that every joule of radiation that strikes the object is converted into thermal energy (none is reflected). The subsequently emitted radiation comes out in a spectrum distributed in a specific way that depends upon the body's temperature.

Physicists in the 19th century devised a clever way to construct an object that imitates very closely the properties of a blackbody. They cut out a cavity from a block of material and left a small hole as an entrance, so that any radiation that enters that hole would be partially absorbed (turned into thermal energy), and partially reflected, but the reflected part is not likely to return out the hole. Instead, it will strike another interior wall of the cavity, again being partially absorbed. By the time the radiation has reflected enough times to accidentally return out the hole, pretty much all of its energy has been converted to thermal. Therefore, for the region where the hole exists, the cavity behaves exactly like a blackbody – all of the incoming radiation is converted to thermal energy, and the only energy that comes out of that hole is due to thermal motions of the charged particles in the material.

Figure 5.1.2 – Radiative Cavity as a Blackbody



In the late 1800's, experiments were performed with these makeshift blackbodies. Note that since the charged particles are moving thermally, their motions can have any of a wide range of frequencies (and therefore so can the EM waves they emit). Physicists were interested in explaining how the outgoing energy was distributed among the emerging EM waves of various wavelengths, and how this distribution depended upon the temperature of the material. The theory that attempts to explain this frequency distribution is steeped in statistical physics (a field of physics that handles issues like the randomness exhibited by the vibrational frequencies of thermally-excited electrons in our blackbody material) that we will not cover in detail in this course. We'll say something about the shape of this distribution shortly, but it should first be noted that the range of frequencies that represent the biggest share of the energy emitted depends upon the temperature of the blackbody. The lower its temperature, the lower the range of peak frequencies lies.

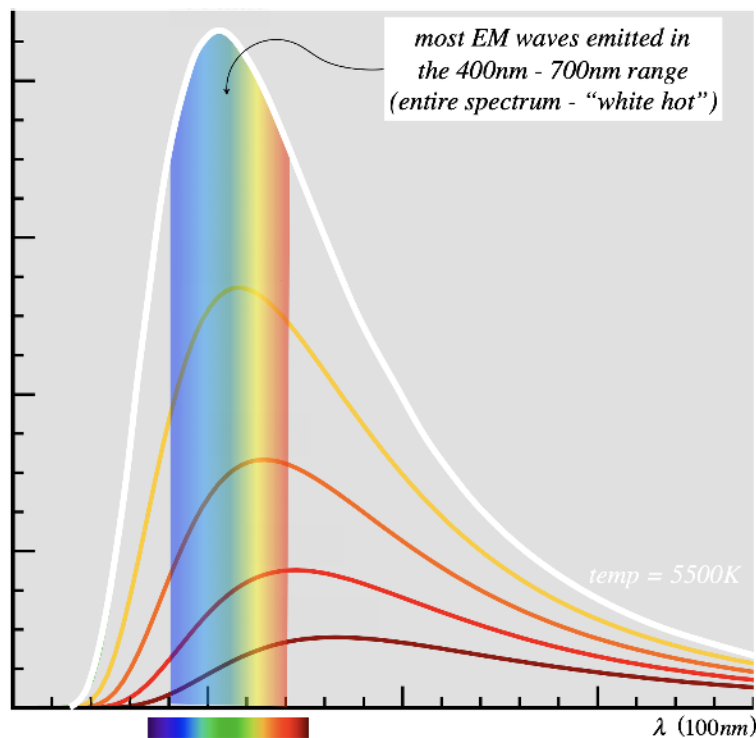
We are already somewhat familiar with this fact. We know that when we heat a piece of metal, we can first start to feel heat radiating from it (infrared radiation – lower frequency than visible light), then when it gets hotter it starts to “glow” red, then hotter still, red starts to be accompanied by other, shorter wavelengths, making it “white hot,” and so on.

Figure 5.1.3 – Glowing Hot Metal



So we would (correctly) guess that the most populated wavelength emitted by a blackbody gets shorter as the temperature rises. If we were to plot the distribution of energy as a function of wavelength, we would expect it to peak at the wavelength ordained by the temperature to be the most common, and re-plotting this graph for a new temperature would shift the peak. The figure below shows several plots of intensity (brightness) of light emitted by a blackbody, as a function of wavelength. Each plot represents the distribution for a blackbody of a different temperature. The range of wavelengths on the horizontal axis that corresponds to the visible spectrum is shown.

Figure 5.1.4 – Blackbody Radiation Curve



The lowest curve on in this figure corresponds to the lowest temperature of the group, and the highest (white) curve corresponds to the hottest (5500K). The temperature for the white curve is the surface temperature of the Sun, and as you can see, this corresponds to peaks emission in the visible spectrum. This is no coincidence, as what is "visible" is defined by what creatures on Earth are able to perceive with their senses, and these senses evolved in the presence of this very light.

Let's say a few words about the mathematical meaning of the curves shown in the graph above. Above we stated that this was a plot of intensity vs. frequency, but that is not quite right. One cannot attribute a brightness to a single, precise wavelength, as the total light energy is finite and it is distributed throughout an infinite number of wavelengths. Each wavelength is therefore responsible for an infinitesimal amount of energy, and therefore has zero intensity. What this curve instead shows is a *spectral density* of light intensity. "Spectral" loosely means "as a function of wavelength" (or equivalently, frequency). With any density, we integrate it over some range to get a total quantity. For example, to get total electric charge in a line segment of charge, we integrate the linear charge density over the range of that segment:

$$Q \text{ (in segment between } x = a \text{ and } x = b) = \int_a^b \lambda(x) dx \quad (5.1.1)$$

Well, the curves above are plots of spectral density Ψ , which means that the "a to b" range is between wavelengths, and the stuff that is added together is light intensity. So let's suppose that we measure the brightness of light coming from a blackbody of a given temperature (which defines one of those curves). Suppose further that we filter the contributions to this brightness, so that we only include light possessing wavelengths in the range λ_a to λ_b . This total brightness is the area under the curve above, integrated from λ_a to λ_b :

$$I_{ab} = \int_{\lambda_a}^{\lambda_b} \Psi(\lambda, T) d\lambda \quad (5.1.2)$$

Back in Physics 9B, we saw that intensity for 3-dimensional waves is power (watts) per area (square meter). This quantity Ψ is the density of this quantity within a wavelength range. The spectral density is a function of both wavelength (the horizontal axis) and the temperature (the five curves shown in the diagram are all different temperatures).

Wien, Rayleigh-Jeans, and the "Ultraviolet Catastrophe"

A fellow named Wilhelm Wien used principles of thermodynamics to show that the relationship between the wavelength with the maximum contribution to the energy output and the temperature is in fact inverse. That is:

$$\lambda (\text{maximum emission}) \propto \frac{1}{T} \quad (5.1.3)$$

As you can see in the series of diagrams for various temperatures, as the temperature goes up, the wavelength at which the peak occurs goes down. Wien goes further, and approximates this curve with a pure guess regarding the function that defines it. He does so with a couple of adjustable constants that can be determined from experiments, but his function breaks down a bit at long wavelengths.

Perhaps a few words here are appropriate regarding techniques of scientific exploration. What Wien did in creating his radiation curve is essentially to reverse-engineer a branch of physics known as “phenomenology.” In physics, phenomenology is the field where the theory is bridged to experiment – coming up with ways to test theories in a lab. The reverse-engineering aspect of this discipline consists of admitting for the moment that we don’t understand the exact mechanism involved in an observation (i.e. we don’t have a theory, either because we are missing a piece to the puzzle, or the problem is too complicated to create an adequate model), but we can still create a mathematical model that reasonably predicts what we experimentally observe. It is clearly not an “answer,” and is obviously far less satisfying than a theory from first principles, but it can lead to progress nonetheless.

A couple of British physicists named Rayleigh and Jeans set out to derive this radiation curve from first principles. There are two key elements to this calculation. The first is that the energy of any given EM wave depends only upon the amplitude, not the frequency. That is, waves of blue light carry the same energy as waves of red light if their amplitudes are the same – a standard principle from wave mechanics that we learned in Physics 9B. And second, we need a good model for determining the expected *populations* of waves of various wavelengths, given a certain temperature. It turns out that the (at the time) recently-developed field of thermodynamics had an answer to this second element using something called the Boltzmann distribution and equipartition of energy. But when Rayleigh and Jeans applied this using a standing wave model within the blackbody cavity, they found that the fraction of the total energy that goes to the shorter wavelengths does not rapidly taper-down as the experiments indicated (i.e. as shown in the graphs we have seen), but rather continues to grow! The lack of agreement between the calculation of Rayleigh and Jeans and the sharp drop of the blackbody radiation curve (which occurs at wavelengths just shorter than visible violet light) became known as the “ultraviolet catastrophe.”

Planck's Solution

A German physicist named Max Planck spent several years at the end of the 19th century working on the problem of the blackbody radiation curve. He first found that he could get the right formula by making a small tweak to Wien’s “pure guess” function, and then tried to reverse-engineer it back to the first principles of Rayleigh and Jeans. He found he couldn’t do that, so there had to be something wrong with the fundamentals behind the Rayleigh-Jeans solution.

He finally hit upon a solution right at the century’s turn, but it involved changing what seemed to be an immutable assumption. Rather than assume that the energy density of an EM wave is independent of its frequency, he found that assuming that EM waves occur in individual packets of energy proportional to their frequencies gave the correct blackbody radiation law. In particular, he found that if he assumed that if an individual EM wave of frequency f carried an energy of hf (where h is a constant, and of course $\lambda f = c$), then using Boltzmann’s statistics, the equation of the spectral energy density curve would come out to what he had already determined must be correct:

$$\Psi(\lambda) = \left(\frac{8\pi hc}{\lambda^5} \right) \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1} \quad (5.1.4)$$

It’s worth emphasizing how radical this was: Planck knew that the curve expressed by the function above matched the experimental data, and he was able to derive it from physical principles, but the idea that the energy of an EM wave could be proportional to the frequency was so far outside the understanding of the time that Planck was certain there was a problem somewhere that would eventually be worked-out. He published the result nonetheless.

It should be clear intuitively why this assumption would help avert the ultraviolet catastrophe. With the Rayleigh-Jeans assumption, there was no “energy penalty” for producing EM waves in higher and higher frequencies (shorter and shorter wavelengths), and since Boltzmann statistics allowed for *more* EM waves of higher frequencies than of lower frequencies, the equipartition theorem assured that a disproportionate amount of energy would go to those high frequency waves. But with the

assumption of Planck, the higher the frequency we choose, the greater slice out of the total energy pie must be taken, leaving less remaining pie for other high frequency waves. This changes the statistics, weighting them against the higher frequencies, and causing the max value to hit a peak. Put another way, eventually the frequency will be so large (wavelength so small), that there simply isn't enough energy in the whole blackbody to equal hf , so that wave will never be created. The constant h is now referred to as *Planck's constant*, and has the value:

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

The idea that light energy is proportional to frequency obviously begged to be explained, but this would have to wait for awhile. In the meantime, there was other evidence emerging that it is true. We will look at a couple more of these clues.

This page titled [5.1: Blackbody Radiation](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

5.2: The Photoelectric Effect

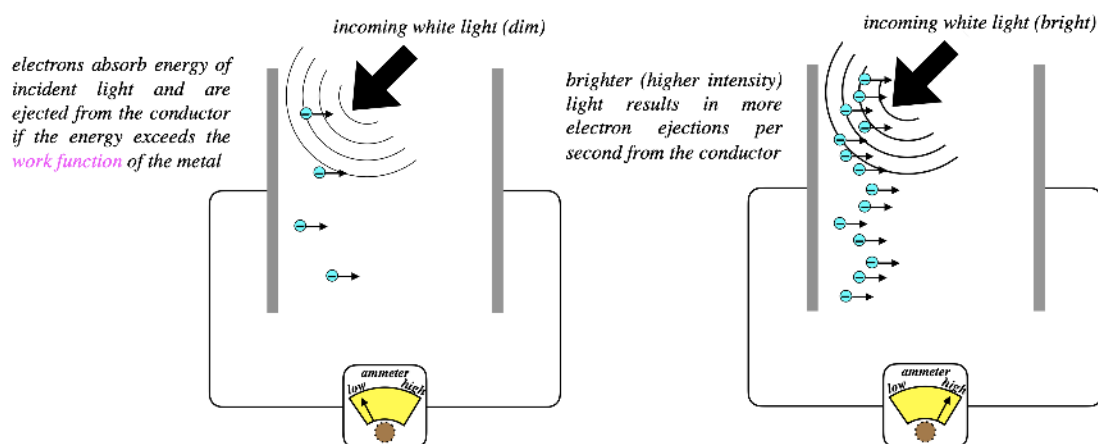
Light Interacting with Conductors

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 second 1905 paper by Einstein, explaining a phenomenon relating to light striking a conducting surface.

When we studied EM in Physics 9C, we always assumed that charges on the surfaces of conductors remained on those conductors. But this belied the fact that we would 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 from 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 5.2.1 - Photoelectric Effect (Unsurprising)



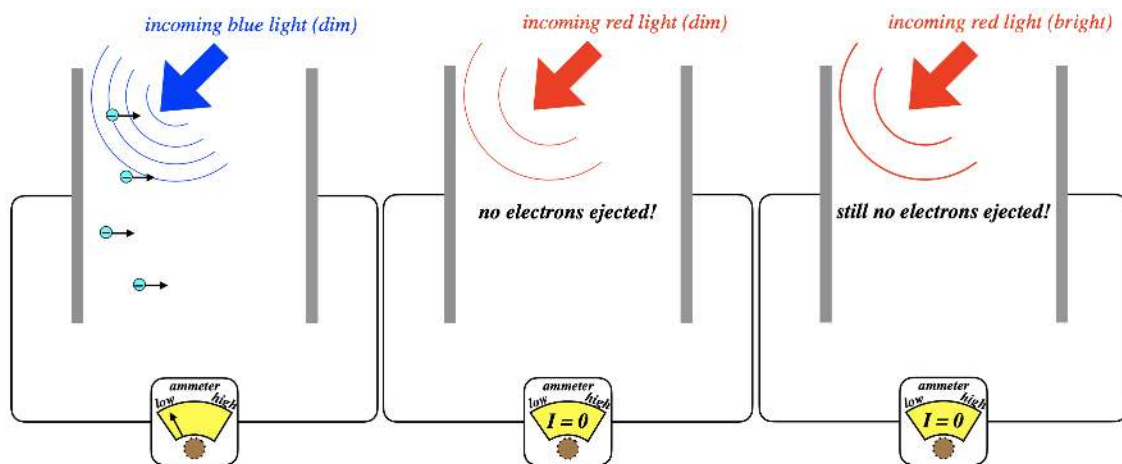
Digging Deeper

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 5.2.2 - Frequency Dependence of Photoelectric Effect



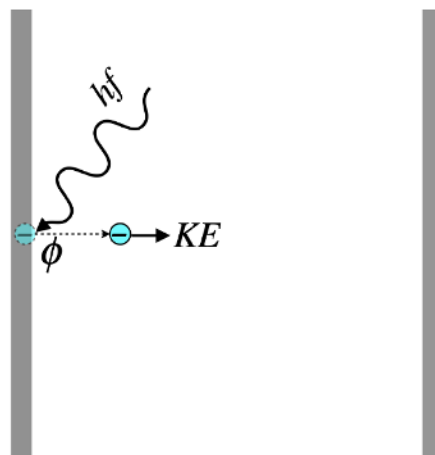
The blue light ejects electrons even when dim, as the white light did, but dim red light does not. This tells us that it was the blue part of the spectrum that was ejecting the electrons when the white light was shone on the metal earlier. While this result is peculiar from our standard understanding 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.

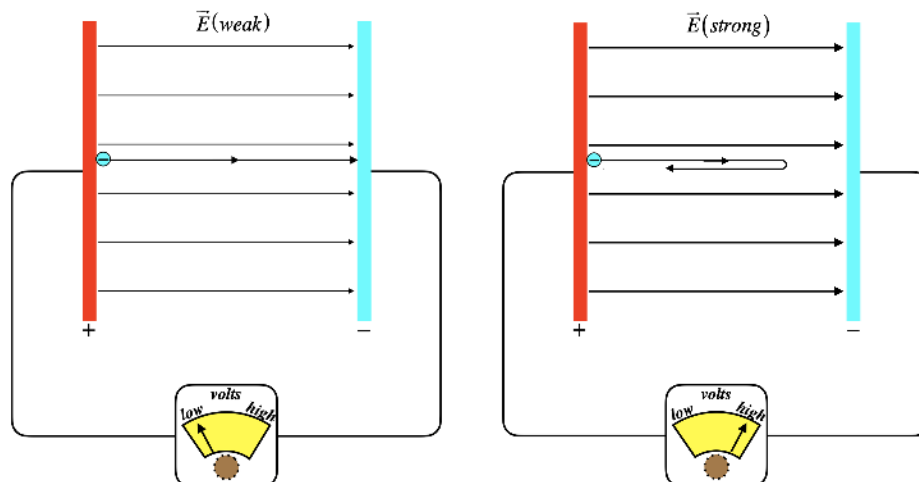
This answers the effect of frequency, but what about the second "digging deeper" question regarding the energy of the electrons that are ejected? Einstein's solution gives us that answer as well. Applying conservation of energy to this process gives us immediately what we seek:

Figure 5.2.3 - 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 measured something called the *stopping potential*, which works like this:

Figure 5.2.4 - 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:

$$e\Delta V_{\text{stopping}} = KE_{\text{max}} = hf - \phi \quad (5.2.1)$$

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)."

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.

This page titled [5.2: The Photoelectric Effect](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

5.3: Compton Effect

Scattering

In Physics 9A, we spend some time talking about [collisions in two dimensions](#), such as two billiard balls. Since then, 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 what we found in [Equation 4.2.7](#):

$$E_{\text{photon}} = pc \quad (5.3.1)$$

It can be shown that it is not possible for both energy and momentum to be conserved in the case of a free particle, but if the photon scatters off the electron, these conservation laws can be obeyed. Of course, the photon will in general give some of its energy to the particle, so the resulting scattered photon is not really the same photon as before (it has a different frequency!). A before/after picture of the event is helpful:

Figure 5.3.1 – 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:

$$\begin{aligned} x\text{-direction:} \quad p_{in} &= p_{out} \cos \theta + \gamma_u m u \cos \phi \\ y\text{-direction:} \quad 0 &= p_{out} \sin \theta - \gamma_u m u \sin \phi \end{aligned} \quad (5.3.2)$$

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

$$p_{in}c + mc^2 = p_{out}c + \gamma_u mc^2 \quad (5.3.3)$$

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

$$\frac{1}{\lambda_{in}} - \frac{1}{\lambda_{out}} = \frac{1}{m\lambda_C} (1 - \cos \theta) \quad (5.3.4)$$

$$p_{out} = p_{in} + mc$$

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:

$$E_{\text{photon}} = pc = hf = h \left(\frac{c}{\lambda} \right) \Rightarrow p = \frac{h}{\lambda} \quad (5.3.5)$$

Putting this in above gives:

$$\lambda_{out} - \lambda_{in} = \frac{h}{mc} (1 - \cos \theta) \quad (5.3.6)$$

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, $\cos \theta = \cos 180^\circ = -1$, 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.

This page titled [5.3: Compton Effect](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

5.4: Double-Slit Experiment

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 the experiments of 9B (polaroids, double slit interference, etc.) then it 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!! 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

Perhaps nothing demonstrates the wave nature of light better than the phenomenon of double slit interference. Let's start with a quick refresher...

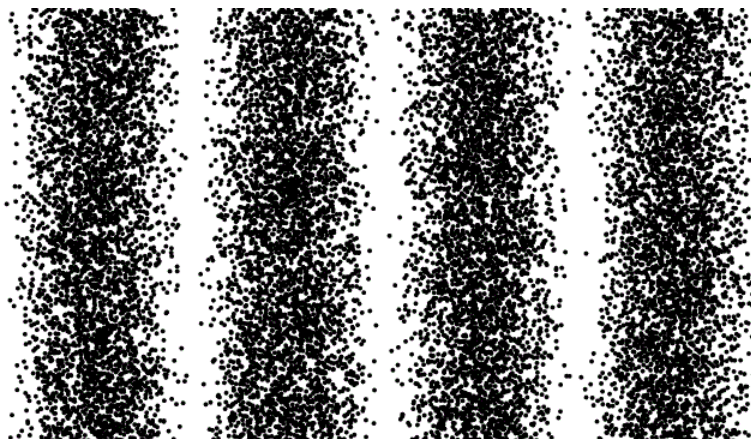
Light (a wave) arrives in phase at two slits, and part of the wave originated from each slit. The light (a wave) from each slit source spreads out to all parts of the screen and at each point on the screen the two waves arrive at some relative phase. If the phases are different by an even integer times π (because of the path length difference), then the light (a wave) interferes constructively, doubling the amplitude and quadrupling the intensity. If the phases are different by an odd integer times π , then the light (a wave) interferes destructively, resulting in zero intensity.

This seems like a phenomenon that can't possibly accommodate photons as particles. But our study of the photoelectric effect gives us a way to test this possibility. In the photoelectric effect, we said that the overall energy that hits a metal plate has two factors that play a role: The energy of each photon, which is proportional to that photon's frequency (the color of the light), and the total number of photons landing on the plate per second (the intensity, or brightness of the light). Suppose we greatly reduce the intensity of the light, to the point where only one photon is fired through a double-slit apparatus per hour. Only a wave can pass through both slits at once and interfere with itself on the other side. A particle has to pick one slit or the other to go through. So by reducing the intensity to one photon per hour, we should not see a light interference pattern. But if the light is actually a wave, we *should* see an interference pattern, albeit an extremely dim one.

So what happens when we do this? Well, in fact we *do* see the one-photon-per-hour scheme demonstrate that photons are particles. Every time a photon is fired through the double-slit, a single dot appears on the screen behind the slit, *not* a continuous distribution of light brightness displaying the interference pattern! So that settles it, light is a particle, and apparently when many particles are fired together, they affect each other's paths (maybe by bumping into each other) in a manner that they replicate a wave interference pattern, tricking us into thinking that light is a wave.

Well this interpretation certainly seems reasonable until someone accidentally leaves the machine on during the month-long winter break. When they come back, they see the dots on the screen where the hundreds of photons landed after passing through the double slit. They 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-per-hour rate.

Figure 5.4.1 – 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, and 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. Indeed the use of the very word "quantum" goes to the heart of this. 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".

This page titled [5.4: Double-Slit Experiment](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

5.5: 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:

$$E^2 = p^2 c^2 + m^2 c^4 \Rightarrow E = pc \quad (5.5.1)$$

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:

$$p = \frac{E}{c} = \frac{hf}{c} = \frac{h}{\lambda} \quad (5.5.2)$$

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:

$$d \sin \theta = \lambda, \quad 2\lambda, \quad 3\lambda, \dots \quad \text{where: } \lambda = \frac{h}{p_e} = \frac{h}{m_e v} = \frac{h}{\sqrt{2mE}} \quad (5.5.3)$$

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 $\sin \theta \leq 1$:

$$d \sin \theta = m\lambda \Rightarrow m_{max} \leq \frac{d \sin \theta_{max}}{\lambda} = \frac{d}{\lambda} \quad (5.5.4)$$

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:

$$E = pc = \frac{hc}{\lambda} = \frac{(6.63 \times 10^{-34} Js) (3 \times 10^8 \frac{m}{s})}{10^{-9} m} = 2.0 \times 10^{-16} J \quad (5.5.5)$$

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

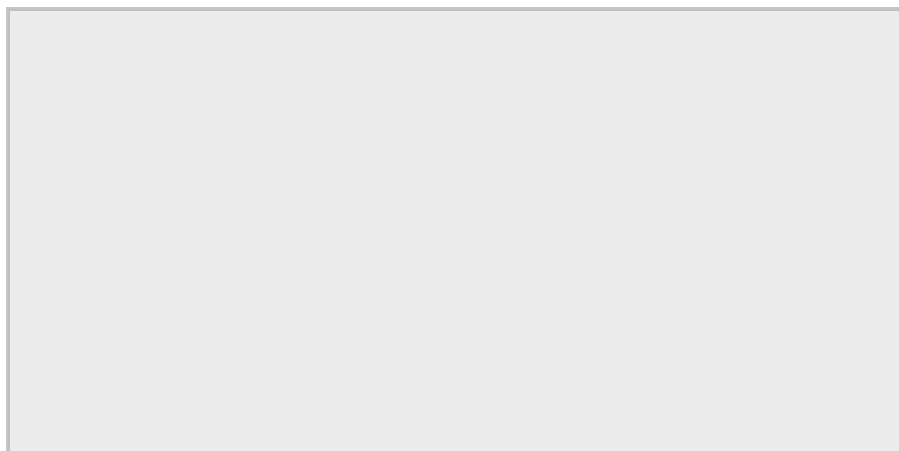
$$E = \frac{p^2}{2m_e} = \frac{h^2}{2\lambda^2 m_e} = \frac{(6.63 \times 10^{-34} Js)^2}{2 (10^{-9} m)^2 (9.11 \times 10^{-31} kg)} = 2.4 \times 10^{-19} J \quad (5.5.6)$$

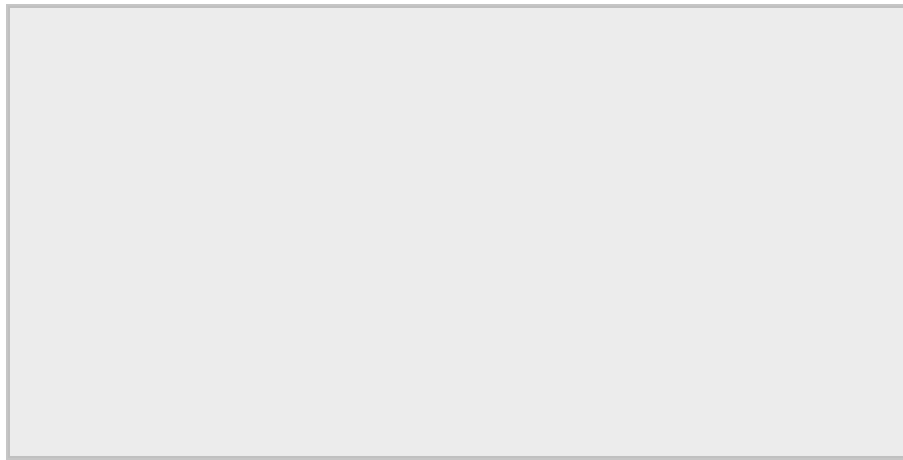
Almost 1000 times as much energy is needed for x-rays than is needed for an electron beam, to achieve the same resolution.

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.

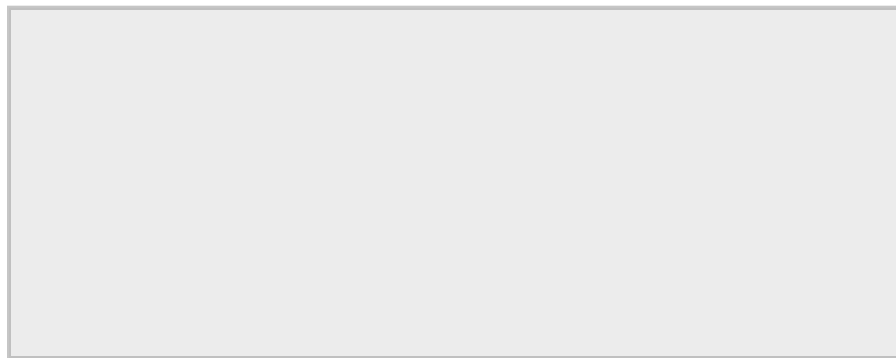
Figure 5.5.1 – Electrons Through a Double Slit with One Slit Blocked





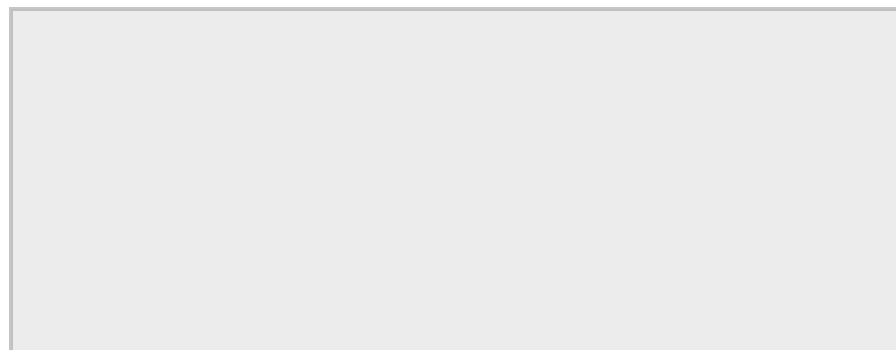
Given these results, the natural prediction of what happens when both slits are open is that we see both of the patterns we saw previously, at the same time. All the electrons that pass through the top slit should end up across from it, and all of those that pass through the bottom slit end up across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

Figure 5.5.2 – 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 5.5.3 – 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, its 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.

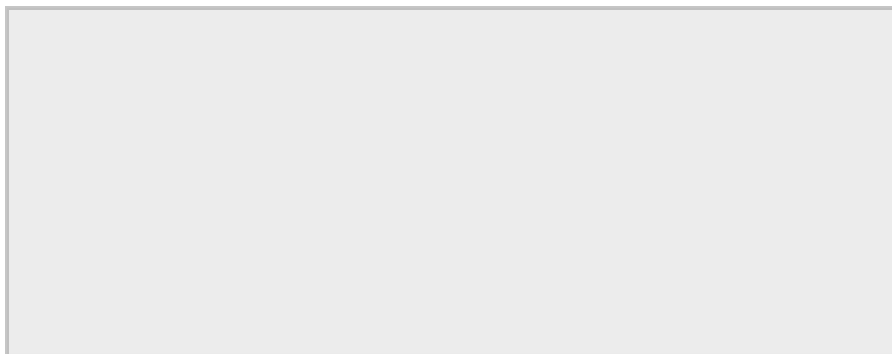
This page titled [5.5: Matter Has Wave Properties, Too!](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

5.6: Summarizing this Wave/Particle Mess

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.

Suppose we have a double slit with an unknown separation and want to determine that separation (we performed this exact experiment in 9B lab!). The way to do it is to send some light through it, observe the interference, and do some math. We have the relation between wavelength λ , slit separation d , and angle θ that the bright fringe is deflected to:

$$d \sin \theta_m = m\lambda \quad (5.6.1)$$

We can solve this for d , and voilà, we know how far apart the slits are. What happens if we use light with a very long wavelength? Well, the quantity $\sin \theta_m$ can never be larger than 1, so for a sufficiently long wavelength of light (namely, longer than d), there is no interference pattern (no dark fringes at all, just one bright fringe at the center). That is, there are no θ 's to plug in. We can't *resolve* the two slits – we can't draw any conclusions about their separation, or even tell if there *are* two slits! As a general rule of thumb, we can say that in order to resolve two positions like the positions of these two slits, we need to use a wave that has a wavelength shorter than the dimension we are trying to resolve. So if we want to watch the electrons so that we can see which slit they go through, we need to use light with a wavelength smaller than 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 9B, 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 ideal gas tells us exactly its temperature), in quantum physics, we can usually only know *probabilities* of the state's values being measured.

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: We introduce something called a *wave function* $\psi(x)$, which has the following properties:

- carries information about the probability of the particle being measured ("make a dot on a screen") 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.

This page titled [5.6: Summarizing this Wave/Particle Mess](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

CHAPTER OVERVIEW

6: The Universe is Inherently Probabilistic

[6.1: Basics of Probability Theory](#)

[6.2: Continuous Probability Distributions and Probability Density](#)

[6.3: The Uncertainty of Random Outcomes](#)

[6.4: Physical Measurements with Random Outcomes](#)

[6.5: Incompatible Measurements](#)

This page titled [6: The Universe is Inherently Probabilistic](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

6.1: 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 6.1.1 – Probability Distribution for Sum of Two Six-Sided Dice

roll	2	3	4	5	6	7	8	9	10	11	12
probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

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 asymmetrically), 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 6 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:

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{A and B mutually exclusive} \quad (6.1.1)$$

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:

$$P(A \text{ and } B) = P(A) \cdot P(B) \quad A \text{ and } B \text{ independent} \quad (6.1.2)$$

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 die 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:

$$\langle \omega \rangle = P(A) \cdot \omega(A) + P(B) \cdot \omega(B) + \dots \quad (6.1.3)$$

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:

$$\langle \omega \rangle = \left(\frac{1}{4} \right) (+\$4) + \left(\frac{5}{18} \right) (-\$3) + \left(\frac{17}{36} \right) (\$0) = +\frac{1}{6} (\$1) \quad (6.1.4)$$

This page titled [6.1: Basics of Probability Theory](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

6.2: 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: $P(1.2 \leftrightarrow 3.6) = \frac{3.6-1.2}{8} = 0.3$.

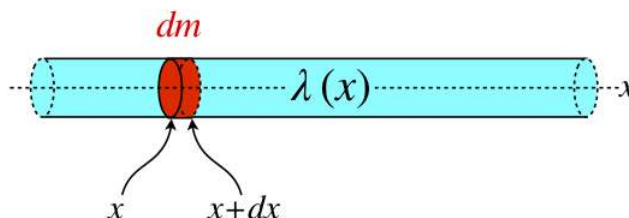
But what if the probability distribution is not uniform, that is, what if the outcomes at some places in the continuum are more probable than others?

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 $\lambda(x)$, the infinitesimal amount of mass found in the tiny slice between positions x and $x + dx$ is given by $dm = \lambda(x) dx$.

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 $\mathcal{P}(x)$ in the same way as we did for mass:

$$dP(x \leftrightarrow x + dx) = \mathcal{P}(x) dx \quad (6.2.1)$$

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:

$$P(x_1 \leftrightarrow x_2) = \int_{x_1}^{x_2} dP = \int_{x_1}^{x_2} \mathcal{P}(x) dx \quad (6.2.2)$$

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 6.1.1](#).

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:

$$1 = \int_{\text{all } x} \mathcal{P}(x) dx \quad (6.2.3)$$

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:

$$1 = \int_{-\infty}^{+\infty} \mathcal{P}(x) dx \quad (6.2.4)$$

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 6.1.3](#) is a sum of the product of the infinitesimal outcome probabilities multiplied by the values for each of the outcomes:

$$\langle \omega \rangle = \int_{-\infty}^{+\infty} \mathcal{P}(x) \omega(x) dx \quad (6.2.5)$$

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 6.2.1

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, $\frac{1}{2}T$):

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

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

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

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:

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

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

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

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:

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

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.

This page titled [6.2: Continuous Probability Distributions and Probability Density](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

6.3: 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 (something we have actually already done many times in Physics 9ABC labs!).

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 in some special way. 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, $\langle\omega\rangle$:

$$\text{separation of the } i^{\text{th}} \text{ result from the mean} = \omega_i - \langle\omega\rangle \quad (6.3.1)$$

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:

$$\langle\omega_i - \langle\omega\rangle\rangle = \frac{1}{N} \sum_{i=1}^N [\omega_i - \langle\omega\rangle] = \frac{1}{N} \sum_{i=1}^N \omega_i - \frac{\langle\omega\rangle}{N} \sum_{i=1}^N (1) = \langle\omega\rangle - \langle\omega\rangle = 0 \quad (6.3.2)$$

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 6.1.3](#).
- Calculate the separation of every outcome from the mean, using [Equation 6.3.1](#).
- 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.

$$\Delta\omega = \sqrt{\frac{1}{N} \sum_{i=1}^N [\omega_i - \langle\omega\rangle]^2} \quad (6.3.3)$$

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:

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

$$\Delta\omega = \sqrt{\langle\omega^2\rangle - \langle\omega\rangle^2} \quad (6.3.4)$$

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 6.2.5). 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 6.3.4.

$$\left. \begin{aligned} \langle\omega\rangle &= \int_{-\infty}^{+\infty} \mathcal{P}(x) \omega(x) dx \\ \langle\omega^2\rangle &= \int_{-\infty}^{+\infty} \mathcal{P}(x) [\omega(x)]^2 dx \end{aligned} \right\} \Rightarrow \Delta\omega = \sqrt{\langle\omega^2\rangle - \langle\omega\rangle^2} \quad (6.3.5)$$

Example 6.3.2

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:

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

This gives us the uncertainty of x :

$$\Delta x = \frac{1}{\sqrt{2}} x_o$$

This page titled 6.3: The Uncertainty of Random Outcomes is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Tom Weideman.

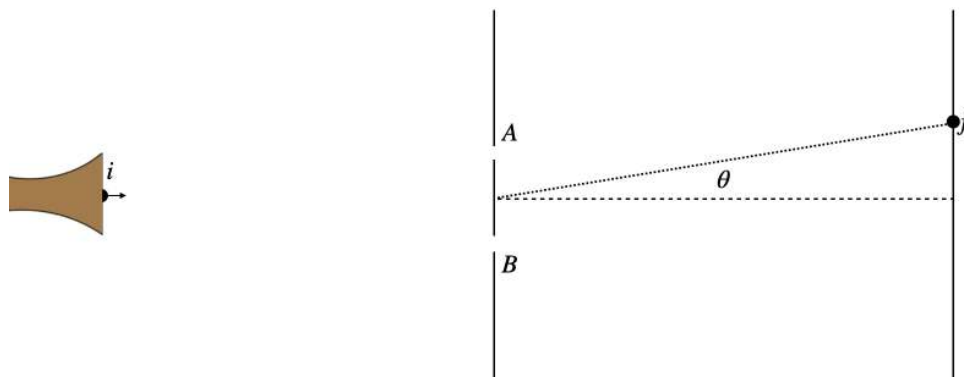
6.4: Physical Measurements with Random Outcomes

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 6.1.1](#)), 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 :

$$P(i \rightarrow f) = P(i \rightarrow A \rightarrow f) + P(i \rightarrow B \rightarrow f) = P_{\text{through } A}(\theta) + P_{\text{through } B}(\theta) \quad (6.4.1)$$

But if we close slit B , the probability $P(i \rightarrow A \rightarrow f)$ 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 5.5.2](#).

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 ' $\psi_A(\theta)$ ' and the wave that goes through slit B ' $\psi_B(\theta)$ ', then the total wave is:

$$\psi_{tot}(\theta) = \psi_A(\theta) + \psi_B(\theta) \quad (6.4.2)$$

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):

$$\mathcal{P}(\theta) = |\psi_{tot}(\theta)|^2 = |\psi_A(\theta) + \psi_B(\theta)|^2 = |\psi_A(\theta)|^2 + |\psi_B(\theta)|^2 + 2\psi_A^*(\theta)\psi_B(\theta) \quad (6.4.3)$$

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*:

$$\psi_{tot}(x) = \psi_1(x) + \psi_2(x) + \dots \quad (6.4.4)$$

- Square the probability amplitude to get the *probability density*:

$$\mathcal{P}(x) = |\psi_{tot}(x)|^2 \quad (6.4.5)$$

- Use the probability density over a range of values to obtain the *probability* that a value lands in a range:

$$P(x \leftrightarrow dx) = \mathcal{P}(x) dx = |\psi_{tot}(x)|^2 dx \quad (6.4.6)$$

- Use the probability density to compute expectation values and uncertainties of physically-measurable quantities:

$$\langle \omega \rangle = \int_{-\infty}^{+\infty} \omega(x) |\psi_{tot}(x)|^2 dx \quad (6.4.7)$$

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 Physics 9B, a solution to the wave equation can be linear combination of many waves, which can all have different wavelengths (this was the source of [beats for sound waves](#)). 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. 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!

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.

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.

This page titled [6.4: Physical Measurements with Random Outcomes](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

6.5: 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 a function familiar from Physics 9B:

$$\Psi(x, t) = A \cos\left(\frac{2\pi}{\lambda}x - 2\pi ft\right) \quad (6.5.1)$$

Let's simplify this discussion by looking at the wave only at time $t = 0$:

$$\psi(x) = \Psi(x, 0) = A \cos\left(\frac{2\pi}{\lambda}x\right) \quad (6.5.2)$$

If we form a probability density from this probability amplitude, we get:

$$\mathcal{P}(x) = |\psi(x)|^2 = A^2 \cos^2\left(\frac{2\pi}{\lambda}x\right) \quad (6.5.3)$$

Okay, let's normalize our probability density (i.e. find the value of A):

$$1 = \int_{\text{all } x} \mathcal{P}(x) dx = \int_{-\infty}^{+\infty} A^2 \cos^2\left(\frac{2\pi}{\lambda}x\right) dx \quad (6.5.4)$$

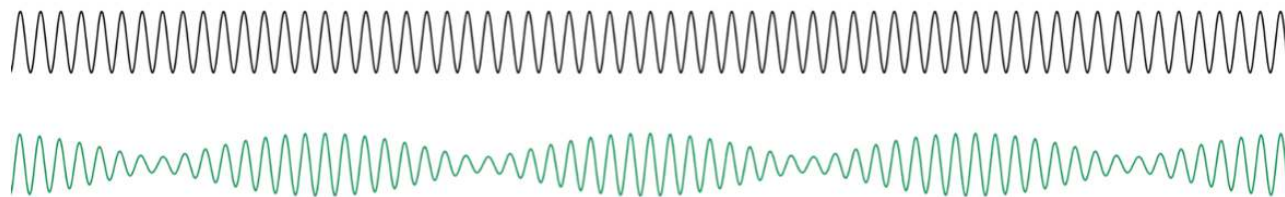
Uh-oh. We have a problem. This integral blows up, making $A = 0$, and $\mathcal{P} \equiv 0$. This makes no sense, what is going on here?

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 6.5.1 – 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):

$$\psi(x) = A \cos kx$$

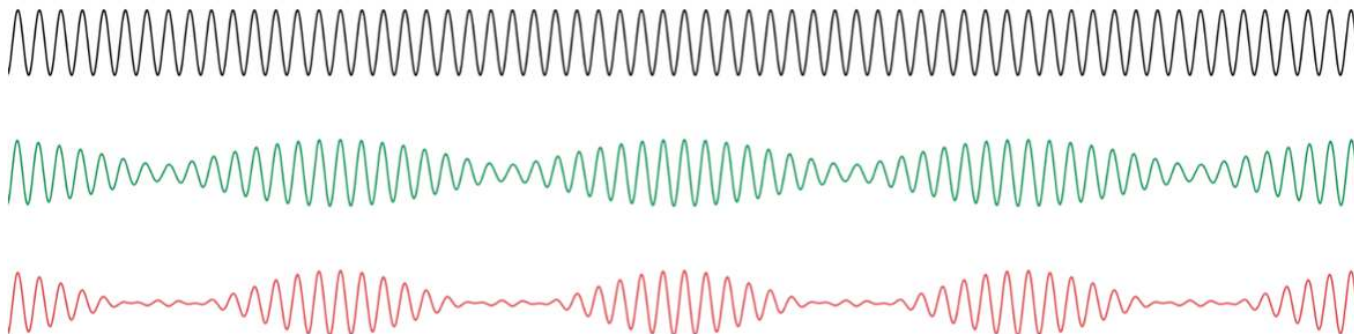
$$\psi(x) = A \cos kx + 0.6A \cos(k + 4\delta k)x$$

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 6.5.2 – Superposition of Three Plane Waves

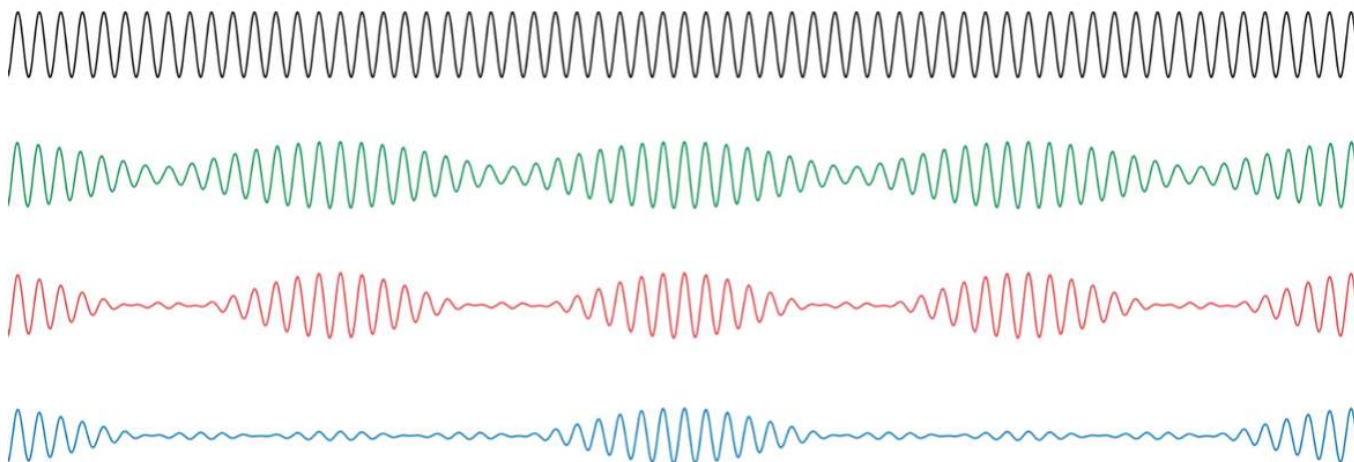


All we did here was add a third plane wave with a wave number *below* that of the original wave by the same amount as the second wave's wave number was larger:

$$\psi(x) = A \cos kx + 0.6A \cos(k + 4\delta k)x + 0.6A \cos(k - 4\delta k)x$$

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 6.5.3 – 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

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 we need to use an integral to capture all of the harmonics. Writing a general function as an integral (essentially a sum) of a continuum of harmonic functions is called taking a *Fourier transform*, which looks like this:

$$\psi(x) = \int_{-\infty}^{+\infty} [A(k) \cos kx + iA(k) \sin kx] dk \quad (6.5.5)$$

The function $A(k)$ is the "recipe" for the transform, meaning it tells us how many "tablespoons" of the harmonic wave of wave number $k = \frac{2\pi}{\lambda}$ to mix into the bowl to get the function ψ . We often refer to this function as the *spectral content* of the particle's quantum state. Yes, $A(k)$ 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:

$$\psi(x) = \int_{-\infty}^{+\infty} A(k) e^{ikx} dk \quad (6.5.6)$$

If we have the function and want to know the recipe, we can extract it using a mathematical trick (that we won't go into here):

$$A(k) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \psi(x) e^{-ikx} dx \quad (6.5.7)$$

This is usually referred to as the *inverse Fourier transform*.

This "recipe" is a function of wavelength, and thanks to deBroglie, we know that wavelength is related to momentum. In fact it turns out that this recipe function is itself a wave function. Just as $\psi(x)$ is a probability amplitude for position x , $A(k)$ is a probability amplitude for momentum. It's probably helpful to explicitly show the link between k (wavenumber) and p momentum. They differ only by a constant factor:

$$k = \frac{2\pi}{\lambda}, \quad p = \frac{h}{\lambda} \Rightarrow p = \frac{h}{2\pi} k = \hbar k \quad (6.5.8)$$

Planck's constant often appears in a fraction with 2π in the denominator, so that fraction is given its own symbol, shown above: $\hbar \equiv$ "h-bar".

When treating the function $A(k)$ as a wave function for momentum, it is common to write it differently, so that its nature is more evident:

$$A(k) \rightarrow \phi(p) \quad (6.5.9)$$

The Heisenberg Uncertainty Principle

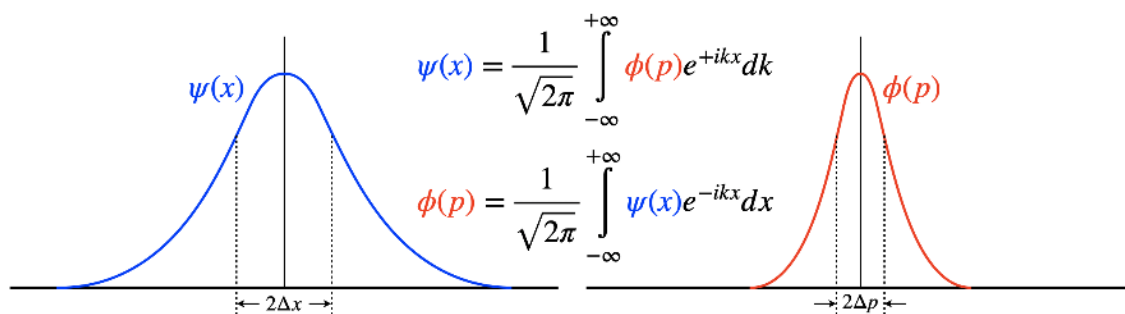
Now that it's quite clear that there is an inverse relationship between the uncertainty of measurements of position and momentum, we can state it formally as was first done by Werner Heisenberg. The specific predicament of a particular particle will define what the uncertainty will be in measuring either 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:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \quad (6.5.10)$$

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).

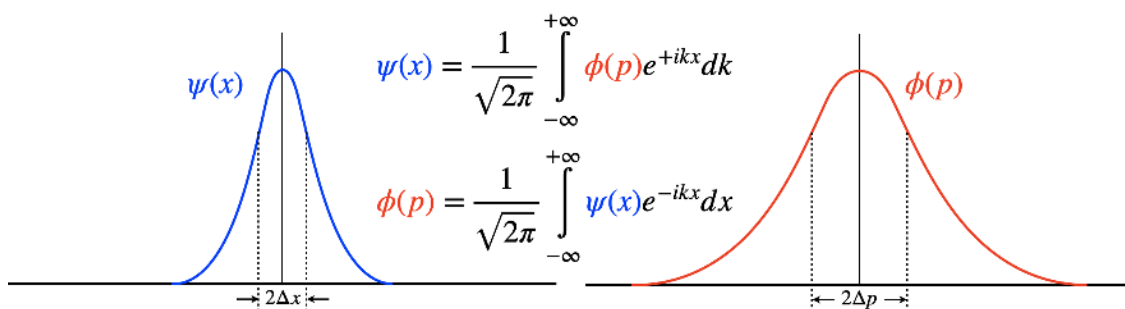
Figure 6.5.4 – Particle Wave Function Expressed in Terms of Position and Momentum



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 6.5.5 – 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](#).

This page titled [6.5: Incompatible Measurements](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

CHAPTER OVERVIEW

7: Matter Waves

[7.1: The Schrödinger Wave Equation](#)

[7.2: States of Definite Energy](#)

[7.3: Operators and Observables](#)

[7.4: Eigenstates and Eigenvalues](#)

This page titled [7: Matter Waves](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

7.1: 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.

$$f(x, t) = A \cos\left(\frac{2\pi}{\lambda}x \pm 2\pi ft\right) \quad (7.1.1)$$

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:

$$\begin{aligned} E = hf \quad \text{and} \quad p = \frac{h}{\lambda} \\ \Rightarrow A \cos\left(\frac{2\pi}{\lambda}x \pm 2\pi ft\right) = A \cos\left(\frac{2\pi p}{h}x \pm \frac{2\pi E}{h}t\right) = A \cos\left(\frac{p}{h}x \pm \frac{E}{h}t\right) \end{aligned} \quad (7.1.2)$$

If we plug this wave function into the wave equation that we know so well, we get confirmation that it works for light:

$$\begin{aligned} \frac{\partial^2}{\partial x^2} A \cos\left(\frac{p}{h}x \pm \frac{E}{h}t\right) &= \frac{1}{c^2} \frac{\partial^2}{\partial t^2} A \cos\left(\frac{p}{h}x \pm \frac{E}{h}t\right) \\ \Rightarrow -\frac{p^2}{h^2} A \cos\left(\frac{p}{h}x \pm \frac{E}{h}t\right) &= -\frac{1}{c^2} \frac{E^2}{h^2} A \cos\left(\frac{p}{h}x \pm \frac{E}{h}t\right) \Rightarrow E = pc \end{aligned} \quad (7.1.3)$$

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:

$$KE = \frac{p^2}{2m} \quad (7.1.4)$$

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:

$$\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi(x, t) = \hbar \frac{\partial}{\partial t} \psi(x, t) \quad (7.1.5)$$

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):

$$\hbar^2 \frac{\partial^2}{\partial x^2} \psi(x, t) = -2m \frac{\partial}{\partial t} \psi(x, t) \quad (7.1.6)$$

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x, t)}{\partial x^2} = i\hbar \frac{\partial \psi(x, t)}{\partial t} \quad (7.1.6)$$

Harmonic (plane wave) solutions to this differential equation look like:

$$\psi(x, t) = \psi_o \left[\cos\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) + i \sin\left(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t\right) \right] \quad (7.1.7)$$

We can shorten the formula for the wave function using the Euler identity:

$$e^{i\theta} = \cos \theta + i \sin \theta \Rightarrow \psi(x, t) = \psi_o e^{i(\frac{p}{\hbar}x \pm \frac{E}{\hbar}t)} \quad (7.1.8)$$

As we saw in [Section 6.5](#), 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):

$$\mathcal{P}(x) = |\psi(x)|^2 = \psi^*(x) \psi(x) \quad (7.1.9)$$

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*:

$$\begin{aligned} \psi_2(x) = e^{i\delta} \psi_1(x) &\Rightarrow |\psi_2(x)|^2 = (e^{i\delta} \psi_1(x))^* (e^{i\delta} \psi_1(x)) \\ &= (e^{-i\delta} \psi_1^*(x)) (e^{i\delta} \psi_1(x)) = e^0 \psi_1^*(x) \psi_1(x) = |\psi_1(x)|^2 \end{aligned} \quad (7.1.10)$$

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.

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 7.1.2](#)) 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:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x, t)}{\partial x^2} + V(x) \psi(x, t) = i\hbar \frac{\partial \psi(x, t)}{\partial t} \quad (7.1.11)$$

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 $V(x)$.

This page titled [7.1: The Schrödinger Wave Equation](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

7.2: 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. 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.

The trick to this method (called *separation of variables*) is to look only for solutions where the function of two variables can be written as the product of two functions of single variables:

$$\Psi(x, t) = \psi(x) \cdot \tau(t) \quad (7.2.1)$$

Plug this into the Schrödinger equation:

$$\begin{aligned} & -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \Psi(x, t) + V(x) \Psi(x, t) = i\hbar \frac{\partial}{\partial t} \Psi(x, t) \\ \Rightarrow & -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} [\psi(x) \cdot \tau(t)] + V(x) [\psi(x) \cdot \tau(t)] = i\hbar \frac{\partial}{\partial t} [\psi(x) \cdot \tau(t)] \end{aligned} \quad (7.2.2)$$

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:

$$-\frac{\hbar^2}{2m} \psi''(x) \tau(t) + V(x) \psi(x) \tau(t) = i\hbar \psi(x) \tau'(t) \quad (7.2.3)$$

Dividing both sides of the equation by $\psi(x) \tau(t)$ gives:

$$\frac{-\frac{\hbar^2}{2m} \psi''(x) + V(x) \psi(x)}{\psi(x)} = i\hbar \frac{\tau'(t)}{\tau(t)} \quad (7.2.4)$$

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:

$$\frac{-\frac{\hbar^2}{2m} \psi''(x) + V(x) \psi(x)}{\psi(x)} = E = i\hbar \frac{\tau'(t)}{\tau(t)} \quad (7.2.5)$$

$$\Rightarrow -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + V(x) \psi(x) = E \psi(x), \quad i\hbar \frac{d}{dt} \tau(t) = E \tau(t) \quad (7.2.6)$$

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:

$$\tau(t) = \tau_0 e^{-i\omega t}, \quad \text{where: } \omega \equiv \frac{E}{\hbar} \quad (7.2.7)$$

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:

$$\mathcal{P}(x, t) = |\Psi(x, t)|^2 = \psi^*(x) \tau^*(t) \psi(x) \tau(t) = |\psi(x)|^2 |\tau_0|^2 e^{+i\omega t} e^{-i\omega t} = |\tau_0|^2 |\psi(x)|^2 \quad (7.2.8)$$

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:

$$\mathcal{P}(x, t) = |\psi(x)|^2 \quad (7.2.9)$$

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:

$$\begin{aligned} \Psi_{tot}(x, t) &= \Psi_1(x, t) + \Psi_2(x, t) = \psi_1(x) e^{-i\omega_1 t} + \psi_2(x) e^{-i\omega_2 t} \\ &\Rightarrow \left[-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right] \Psi_{tot}(x, t) \\ &= \left[-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right] \left[\psi_1(x) e^{-i\omega_1 t} + \psi_2(x) e^{-i\omega_2 t} \right] \end{aligned} \quad (7.2.10)$$

$$\begin{aligned}
 &= \left[-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right] \psi_1(x) e^{-i\omega_1 t} + \left[-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right] \psi_2(x) e^{-i\omega_2 t} \\
 &= E_1 \psi_1(x) e^{-i\omega_1 t} + E_2 \psi_2(x) e^{-i\omega_2 t}
 \end{aligned}$$

Clearly this does not have the form of a stationary state solution, as the functions of time cannot be made to go away.

This page titled [7.2: States of Definite Energy](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

7.3: 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 is 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 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:

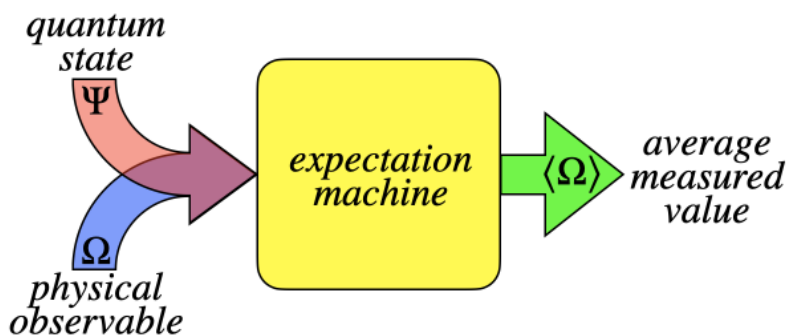
$$\langle x \rangle = \int_{-\infty}^{+\infty} \mathcal{P}(x) x dx = \int_{-\infty}^{+\infty} |\Psi|^2 x dx = \int_{-\infty}^{+\infty} \Psi^* \Psi x dx \quad (7.3.1)$$

If we instead wish 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 7.3.1 – 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 7.3.2 – 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:

$$\Rightarrow \langle x \rangle = \int_{-\infty}^{+\infty} \Psi^* \Psi_{new} dx = \int_{-\infty}^{+\infty} \Psi^* (x \Psi) dx = \int_{-\infty}^{+\infty} x |\Psi|^2 dx = \int_{-\infty}^{+\infty} x \mathcal{P}(x) dx \quad (7.3.2)$$

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:

$$\Rightarrow \langle p \rangle = \int_{-\infty}^{+\infty} \Phi^* \Phi_{new} dk = \int_{-\infty}^{+\infty} \Phi^* (p \Phi) dk = \int_{-\infty}^{+\infty} p |\Phi|^2 dk = \int_{-\infty}^{+\infty} p \mathcal{P}(k) dk \quad (7.3.3)$$

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} :

$$\Psi = \int_{-\infty}^{+\infty} \Phi e^{ikx} dk, \quad \Psi_{new} = \int_{-\infty}^{+\infty} \Phi_{new} e^{ikx} dk = \int_{-\infty}^{+\infty} p \Phi e^{ikx} dk = \int_{-\infty}^{+\infty} \hbar k \Phi e^{ikx} dk \quad (7.3.4)$$

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:

$$\Psi_{new} = -i\hbar \frac{d}{dx} \Psi = -i\hbar \frac{d}{dx} \int_{-\infty}^{+\infty} \Phi e^{ikx} dk = -i\hbar \int_{-\infty}^{+\infty} \Phi \frac{d}{dx} e^{ikx} dk = \int_{-\infty}^{+\infty} \hbar k \Phi e^{ikx} dk \quad (7.3.5)$$

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):

$$\hat{p}_x = -i\hbar \frac{d}{dx} \quad (7.3.6)$$

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:

$$\hat{p}_x \psi(x) = -i\hbar \frac{d}{dx} \psi(x), \quad \hat{p}_x \phi(k) = \hbar k \phi(k) \quad (7.3.7)$$

Similarly, the operator $\hat{x}$ 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 to the reader to determine the operator $\hat{x}$ that acts on $\phi(k)$).

Going back to the original discussion of computing expectation values, we see that we have:

$$\langle p \rangle = \int_{-\infty}^{+\infty} \Psi^* (\hat{p} \Psi) dx = \int_{-\infty}^{+\infty} \Psi^* \left(-i\hbar \frac{d}{dx} \Psi \right) dx \quad (7.3.8)$$

Building More Operators

We can build new operators for other physical observables from $\hat{x}$ and $\hat{p}$. Most notable among these is the kinetic energy operator:

$$\widehat{KE} = \frac{\hat{p}^2}{2m} = \frac{1}{2m} \left(-i\hbar \frac{d}{dx} \right) \left(-i\hbar \frac{d}{dx} \right) = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \quad (7.3.9)$$

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 $\hat{x}$. Together, the operators $\widehat{KE}$ and $\widehat{V}(x)$ account for the total energy, and as a shorthand we sometimes use:

$$\widehat{H} \equiv \widehat{KE} + \widehat{V}(x) \quad (7.3.10)$$

This "total energy operator" is commonly referred to as the **Hamiltonian**. Note that Schrödinger's equation states that the Hamiltonian's actions 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:

$$\widehat{E} \equiv i\hbar \frac{\partial}{\partial t} \quad (7.3.11)$$

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:

$$\Psi_{new} = \hat{x} (\hat{p} \Psi) = x \left(-i\hbar \frac{d}{dx} \Psi \right) = -i\hbar x \frac{d\Psi}{dx} \quad (7.3.12)$$

But if we perform the operation in the other order, we get a different result:

$$\Psi_{new} = \hat{p} (\hat{x} \Psi) = -i\hbar \frac{d}{dx} (x\Psi) = -i\hbar \left(\Psi + x \frac{d\Psi}{dx} \right) \quad (7.3.13)$$

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 achieve the same result when performed in either order, we say that they do not **commute** with each other. Note that any function of $\hat{x}$ (like $\widehat{V}(x)$) will commute with any other function of $\hat{x}$, 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.

This page titled [7.3: Operators and Observables](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

7.4: 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 " $(3.0 + 2.2i) \text{ eV}$ "! 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 $\hat{\Omega}$. Then, using our "expectation machine" from the previous section, we have:

$$\langle \omega \rangle = \omega_1 = \int_{-\infty}^{+\infty} \psi_1^*(x) \left[\hat{\Omega} \psi_1(x) \right] dx \quad (7.4.1)$$

The question is, in what way does the operator $\hat{\Omega}$ 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 :

$$\begin{aligned} \int_{-\infty}^{+\infty} \psi_1^*(x) \left[\hat{\Omega} \psi_1(x) \right] dx &= \int_{-\infty}^{+\infty} \psi_1^*(x) [\omega_1 \psi_1(x)] dx \\ &= \omega_1 \int_{-\infty}^{+\infty} \psi_1^*(x) \psi_1(x) dx = \omega_1 \end{aligned} \quad (7.4.2)$$

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:

$$\hat{\Omega} \psi_i(x) = \omega_i \psi_i(x) \quad (7.4.3)$$

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:

$$\psi(x) = C_1 \psi_1(x) + C_2 \psi_2(x) + \dots = \sum_{\text{all } i} C_i \psi_i(x) \quad (7.4.4)$$

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](#):

$$\int_{-\infty}^{+\infty} \psi_i^*(x) \psi_j(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (7.4.5)$$

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:

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} \psi^*(x) \psi(x) dx \\ &= \int_{-\infty}^{+\infty} [C_1 \psi_1(x) + C_2 \psi_2(x) + \dots]^* [C_1 \psi_1(x) + C_2 \psi_2(x) + \dots] dx \end{aligned} \quad (7.4.6)$$

The integrals of the cross-terms all vanish thanks to the orthogonality condition, leaving:

$$1 = C_1^* C_1 + C_2^* C_2 + \dots = |C_1|^2 + |C_2|^2 + \dots \quad (7.4.7)$$

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:

$$\begin{aligned} \langle \omega \rangle &= \int_{-\infty}^{+\infty} \psi^*(x) [\widehat{\Omega} \psi(x)] dx \\ &= \int_{-\infty}^{+\infty} [C_1^* \psi_1^*(x) + C_2^* \psi_2^*(x) + \dots] [C_1 \widehat{\Omega} \psi_1(x) + C_2 \widehat{\Omega} \psi_2(x) + \dots] dx \\ &= \int_{-\infty}^{+\infty} [C_1^* \psi_1^*(x) + C_2^* \psi_2^*(x) + \dots] [C_1 \omega_1 \psi_1(x) + C_2 \omega_2 \psi_2(x) + \dots] dx \\ &= |C_1|^2 \omega_1 + |C_2|^2 \omega_2 + \dots \\ &= P_1 \omega_1 + P_2 \omega_2 + \dots \end{aligned} \quad (7.4.8)$$

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:

$$\Psi(x, t) = Ae^{i(kx - \omega t)} \quad (7.4.9)$$

We fully expect this to be an eigenstate of momentum, kinetic energy, and total energy, and it is:

$$\begin{aligned} \hat{p}\Psi(x, t) &= -i\hbar \frac{\partial}{\partial x} Ae^{i(kx - \omega t)} = \hbar k \Psi(x, t) \\ \widehat{KE}\Psi(x, t) &= -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} Ae^{i(kx - \omega t)} = \frac{\hbar^2 k^2}{2m} \Psi(x, t) \\ \widehat{E}\Psi(x, t) &= i\hbar \frac{\partial}{\partial t} Ae^{i(kx - \omega t)} = \hbar \omega \Psi(x, t) \end{aligned} \quad (7.4.10)$$

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.

This page titled [7.4: Eigenstates and Eigenvalues](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

CHAPTER OVERVIEW

8: One-Dimensional Models

[8.1: Particle-in-a-Box, Part 1](#)

[8.2: Particle-in-a-Box, Part 2](#)

[8.3: The Finite Square Well](#)

[8.4: Tunneling](#)

[8.5: The Quantum Harmonic Oscillator](#)

[8.6: The Bohr Model of the Hydrogen Atom](#)

This page titled [8: One-Dimensional Models](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

8.1: Particle-in-a-Box, Part 1

Bound States

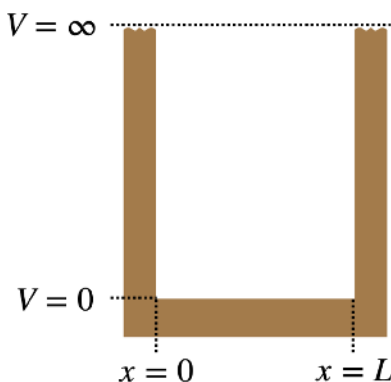
We have discussed at length the case of a free particle, and how we can construct general solutions from plane wave solutions to the Schrödinger equation, but now it's time to have a look at cases where particles are bound to some region by a force. We are staying in one dimension, so this force will 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:

$$V(x) = \begin{cases} 0 & 0 \leq x \leq L \\ \infty & x < 0, x > L \end{cases} \quad (8.1.1)$$

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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + V(x) = E\psi(x) \quad (8.1.2)$$

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):

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + \begin{cases} 0 & \text{(inside well)} \\ \infty & \text{(outside well)} \end{cases} \psi(x) = E\psi(x) \quad (8.1.3)$$

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:

$$\psi(x) = Ae^{+ikx} + Be^{-ikx} \quad (8.1.4)$$

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:

$$\begin{aligned} \psi(0) = 0 &= A + B \\ \psi(L) = 0 &= Ae^{+ikL} + Be^{-ikL} \end{aligned} \quad \Rightarrow \quad e^{+2ikL} = 1 \quad (8.1.5)$$

$$\Rightarrow \quad k_n = \frac{n\pi}{L}, \quad n = 0, 1, 2, \dots$$

We have subscripted the wave number $k \rightarrow k_n$ 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):

$$\psi_n(x) = A(e^{+ik_n x} - e^{-ik_n x}) = 2iA \sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, \dots \quad (8.1.6)$$

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 :

$$\begin{aligned} 1 &= \int_0^L \psi_n^*(x) \psi_n(x) dx = \int_0^L \left[-2iA^* \sin\left(\frac{n\pi x}{L}\right) \right] \left[2iA \sin\left(\frac{n\pi x}{L}\right) \right] dx \\ &= 4|A|^2 \int_0^L \sin^2\left(\frac{n\pi x}{L}\right) dx \end{aligned} \quad (8.1.7)$$

Performing the integral and solving for $|A|^2$ gives:

$$|A|^2 = \frac{1}{2L} \quad (8.1.8)$$

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 $A = \frac{\sqrt{2}}{2}$ we get for the wave function that is the n_{th} harmonic:

$$\psi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, \dots \quad (8.1.9)$$

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:

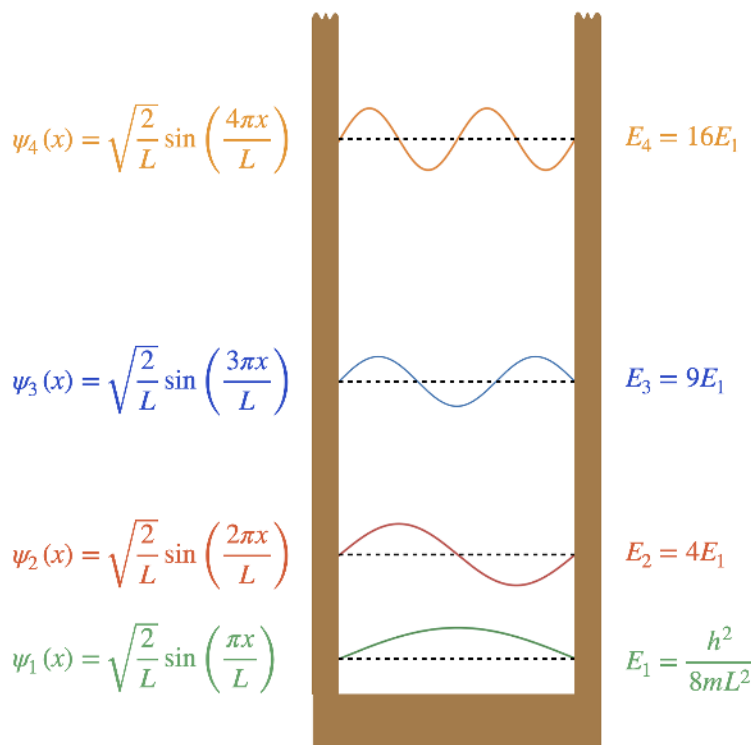
$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \left[\sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \right] = + \frac{\hbar^2 n^2}{2\pi^2 m L^2} \left[\sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \right] \quad (8.1.10)$$

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 :

$$E_n = \frac{\hbar^2 n^2}{2\pi^2 m L^2} = \frac{h^2 n^2}{8mL^2} \quad (8.1.11)$$

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 $|p_n| = \frac{h n}{2L}$.

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 0.1 kg can change its momentum in increments of $\frac{h}{2L}$, so if it is bouncing between walls separated by 20 cm , its "jump" in speed from one level to the next is:

$$v = \frac{p}{m} = \frac{h}{2mL} = \frac{6.63 \times 10^{-34} \text{ Js}}{2 (0.1 \text{ kg}) (0.2 \text{ m})} = 1.66 \times 10^{-32} \frac{\text{m}}{\text{s}} \quad (8.1.12)$$

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*.

This page titled [8.1: Particle-in-a-Box, Part 1](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

8.2: Particle-in-a-Box, Part 2

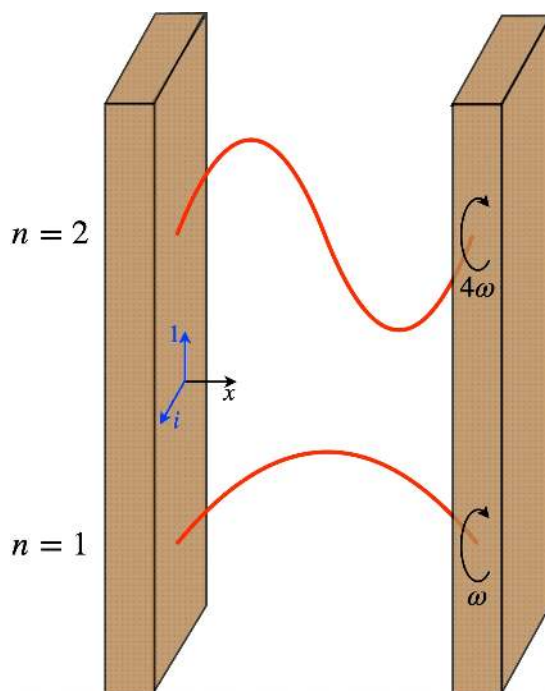
The Time-Dependent Stationary-State Wave Functions

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 the quantum state of one of these energy eigenstates still has a time component to it:

$$\Psi_n(x, t) = \psi_n(x) e^{-i\omega_n t} = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) e^{-i\frac{E_n}{\hbar}t}, \quad E_n = \frac{n^2 \hbar^2}{8mL^2} \quad (8.2.1)$$

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 time-dependent quantum phase as a rotation:

$$e^{-i\omega t} = \cos \omega t - i \sin \omega t \quad (8.2.2)$$

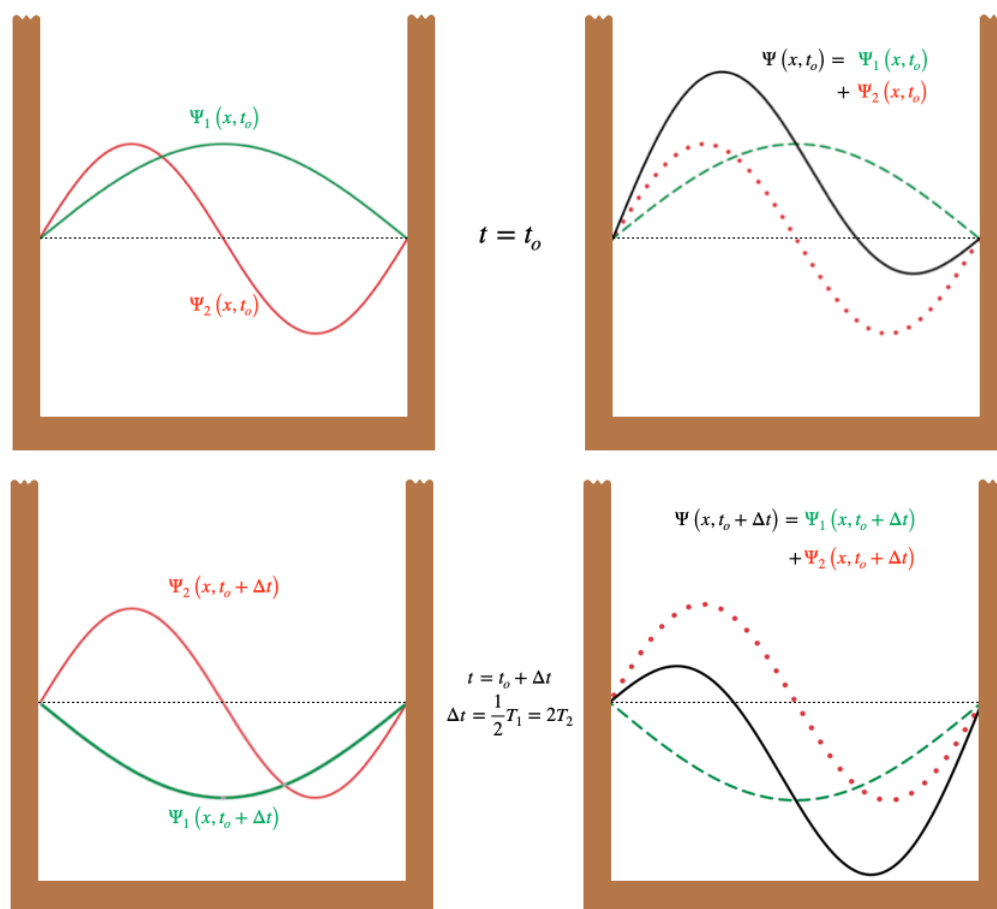
Notice that since $\omega_n \propto E_n \propto n^2$, the $n = 2$ state has 4 times the rotational velocity as the $n = 1$ state.

Given the obvious time-dependence of these solutions, the question arises 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 piece 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:

$$\Psi(x, t) = \sum_{n=1}^{\infty} C_n \psi_n(x) e^{-i\omega_n t} = \sqrt{\frac{2}{L}} \sum_{n=1}^{\infty} C_n \sin\left(\frac{n\pi x}{L}\right) e^{-i\frac{E_n}{\hbar} t} \quad (8.2.3)$$

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:

$$C_n = \int_0^L \psi_n^*(x) \Psi(x, 0) dx \quad (8.2.4)$$

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:

$$hf = \Delta E = E_{n_2} - E_{n_1} \Rightarrow f = \frac{1}{h} \left(\frac{h^2 n_2^2}{8mL^2} - \frac{h^2 n_1^2}{8mL^2} \right) = \frac{h}{8mL^2} (n_2^2 - n_1^2) \quad (8.2.5)$$

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.

This page titled [8.2: Particle-in-a-Box, Part 2](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

8.3: 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.

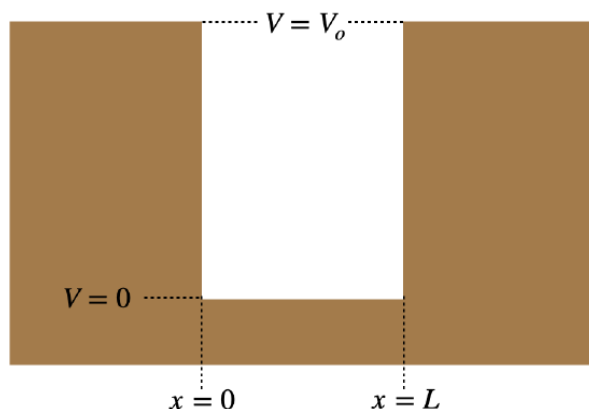
Having said all this, we are interested in the bound states (there is plenty of interesting quantum physics in the unbound case as well, but we will not cover that here). Keep in mind that the ground state is always above the bottom of the well, so if this well is particularly shallow, perhaps we will 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:

$$V(x) = \begin{cases} 0 & 0 \leq x \leq L \\ V_o & x < 0, x > L \end{cases} \quad (8.3.1)$$

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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + \begin{cases} 0 & \text{(inside well)} \\ V_o & \text{(outside well)} \end{cases} \psi(x) = E\psi(x) \quad (8.3.2)$$

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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) = E \psi(x) \Rightarrow \psi_{\text{inside well}}(x) = A e^{+ikx} + B e^{-ikx}, \quad E = \frac{\hbar^2 k^2}{2m} \quad (8.3.3)$$

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:

$$\psi_{\text{inside well}}(x) = A \sin kx + B \cos kx, \quad E = \frac{\hbar^2 k^2}{2m} \quad (8.3.4)$$

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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) = (E - V_o) \psi(x) \quad (8.3.5)$$

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 themselves that this wave function satisfies the differential equation):

$$\psi_{\text{outside well}}(x) = C e^{+\alpha x} + D e^{-\alpha x}, \quad V_o - E = \frac{\hbar^2 \alpha^2}{2m} \quad (8.3.6)$$

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 $x \rightarrow -\infty$. This requires that the coefficient D is identically zero. Similarly, the wave function in the right region will grow without bound as $x \rightarrow +\infty$ unless the constant C is identically zero.

Summarizing what we have so far for the wave function:

$$\psi(x) = \begin{cases} C e^{+\alpha x} & x < 0 \\ A \sin kx + B \cos kx & 0 \leq x \leq L \\ D e^{-\alpha x} & x > L \end{cases} \quad (8.3.7)$$

Energy Spectrum Quantization

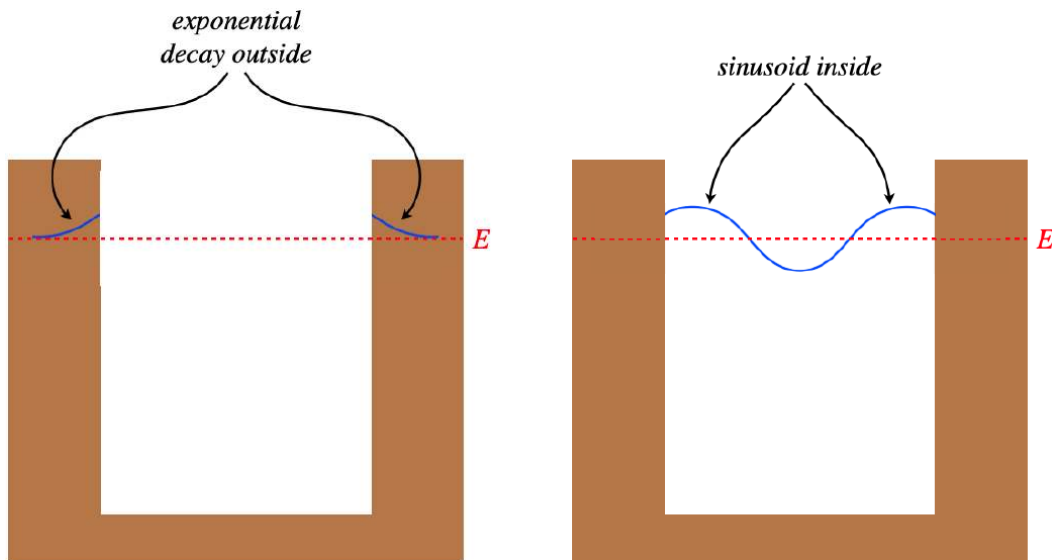
With the presence of a non-zero wave function outside the walls, our task of matching boundary conditions is more daunting than it was for the infinite well. We have two conditions that must hold at the boundaries. The first is that the wave function must be continuous – it is unphysical for the particle 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 these in terms of the known values m (mass of the particle), L (length of the well), and V_0 (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.

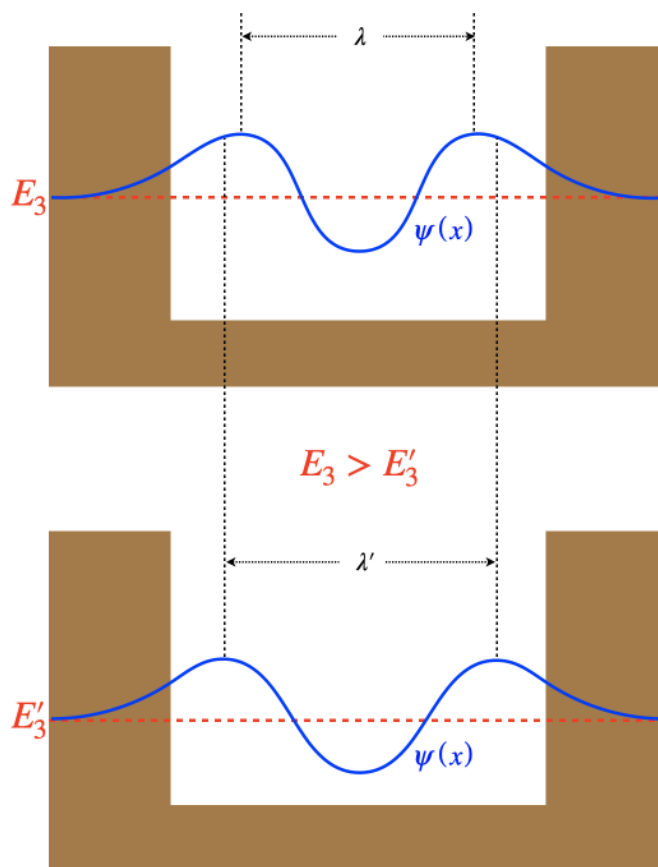
Consider an example of a wave function with three antinodes. What our math so far requires is:

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:

$$\psi(x_o) = A \cos kx_o = B e^{-\alpha x_o} \quad (8.3.8)$$

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:

$$\psi_{new}(x_o) = (A + \Delta A) \cos kx_o = (B + \Delta B) e^{-\alpha x_o} \Rightarrow \Delta A \cos kx_o = \Delta B e^{-\alpha x_o} \quad (8.3.9)$$

The original wave function was constructed to match the derivatives at the boundaries:

$$\psi'(x_o) = -Ak \sin kx_o = -B\alpha e^{-\alpha x_o} \quad (8.3.10)$$

For the derivatives to also match at the boundaries after the change of the values of A and B requires:

$$\begin{aligned} \psi'(x_o) &= -(A + \Delta A) k \sin kx_o = -(B + \Delta B) \alpha e^{-\alpha x_o} \\ \Rightarrow -k\Delta A \sin kx_o &= -\alpha \Delta B e^{-\alpha x_o} \end{aligned} \quad (8.3.11)$$

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 :

$$\frac{-\Delta A k \sin kx_o}{\Delta A \cos kx_o} = \frac{-\alpha \Delta B \alpha e^{-\alpha x_o}}{\Delta B e^{-\alpha x_o}} \Rightarrow k \tan kx_o = \alpha \quad (8.3.12)$$

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** $x = 0$

$$\psi(0) = C e^0 = A \sin 0 + B \cos 0 \Rightarrow C = B$$

$\psi(x)$ **continuous at** $x = L$

$$\psi(L) = D e^{-\alpha L} = A \sin kL + B \cos kL$$

$\frac{d}{dx}\psi(x)$ **continuous at** $x = 0$

$$\psi'(0) = \alpha C e^0 = kA \cos 0 - kB \sin 0 \Rightarrow \alpha C = kA$$

$\frac{d}{dx}\psi(x)$ **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:

$$\tan kL = \frac{2\alpha k}{k^2 - \alpha^2}, \quad k = \frac{\sqrt{2mE}}{\hbar}, \quad \alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar} \quad (8.3.13)$$

One would of course like to solve these for the energy E in terms of m , L , and V_0 , 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.

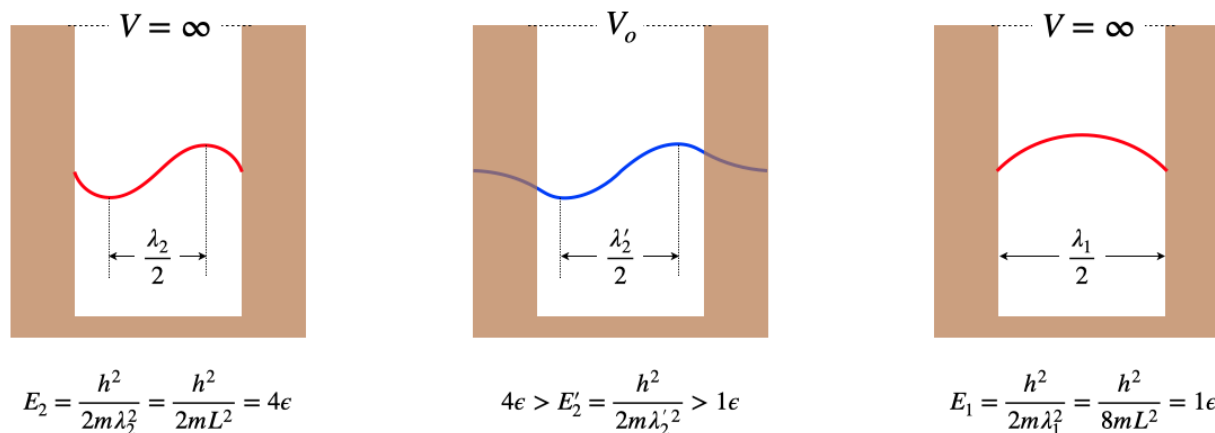
Effects of Varying the Well Depth

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:

$$\epsilon \equiv \frac{h^2}{8mL^2} \quad (8.3.14)$$

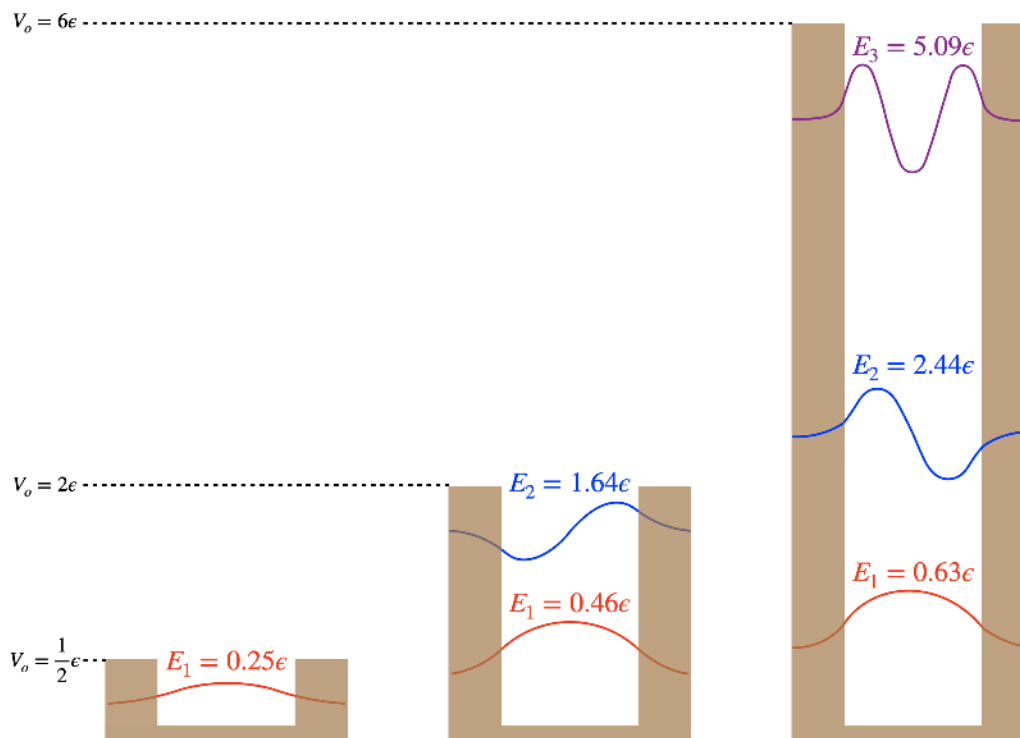
Figure 6.3.5 – Limits of Wavelength for 2 Antinode Wave Function



This clearly limits the number of eigenstates for a given value of V_0 . In particular, there can exist no bound eigenstate with an energy greater than V_0 , 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 $V_0 < n^2\epsilon$.

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 $V_o \rightarrow \infty$ 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 $V_o \rightarrow \infty$, 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](#):

$$\alpha = \frac{\sqrt{2m(V_o - E)}}{\hbar} \quad (8.3.15)$$

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:

$$\lim_{V_o \rightarrow \infty} Ae^{-\alpha|x|} = 0 \quad (8.3.16)$$

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}):

$$\delta \equiv \alpha^{-1} = \frac{\hbar}{\sqrt{2m(V_o - E)}} \quad (8.3.17)$$

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:

$$\psi(x) = Ae^{-\alpha x}, \quad x \geq 0 \quad (8.3.18)$$

Normalizing (so that probabilities and then expectation values work out properly) gives:

$$1 = \int_0^{\infty} |\psi(x)|^2 dx = \int_0^{\infty} |A|^2 e^{-2\alpha x} dx \Rightarrow A = \sqrt{2\alpha} \quad (8.3.19)$$

And now the expectation value of x in this $x \geq 0$ region is:

$$\langle x \rangle = \int_0^{\infty} x [\sqrt{2\alpha} e^{-\alpha x}]^2 dx = 2\alpha \int_0^{\infty} x e^{-2\alpha x} dx = \frac{1}{2\alpha} = \frac{1}{2} \delta \quad (8.3.20)$$

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 difficult, and are left to the reader, but the result is:

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \frac{1}{2} \delta \quad (8.3.21)$$

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:

$$\Delta p \geq \frac{\hbar}{2\Delta x} = \frac{\hbar}{\delta} \quad (8.3.22)$$

The kinetic energy of the particle that has this uncertain momentum is also uncertain. Without working out the math, we can 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 energy, it is a good number to use as a lower limit of this uncertainty. This means that the uncertainty of the kinetic energy is:

$$\Delta KE > \frac{\Delta p^2}{2m} = \frac{\hbar^2}{2m\delta^2} = V_o - E \quad (8.3.23)$$

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.

This page titled [8.3: The Finite Square Well](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

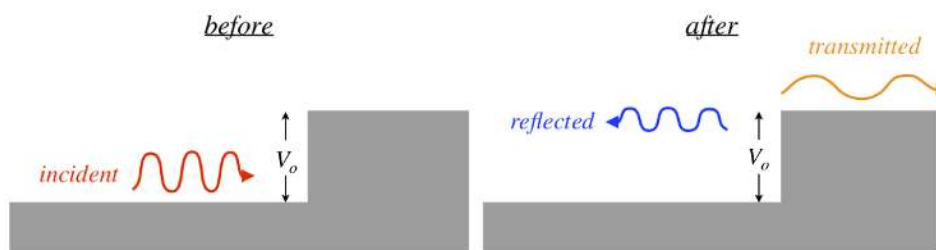
8.4: Tunneling

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:

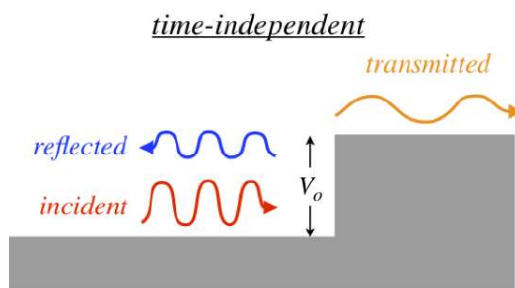
Figure 6.4.1 – Step Potential Before/After Picture



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 there 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:

$$\psi(x) = \begin{cases} \psi_{inc}(x) + \psi_{refl}(x) & x < 0 \\ \psi_{trans}(x) & x > 0 \end{cases} \quad (8.4.1)$$

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:

$$\left. \begin{aligned} \psi_{inc}(x) &= Ae^{+ikx} \\ \psi_{refl}(x) &= Be^{-ikx} \\ \psi_{trans}(x) &= Ce^{+ik'x} \end{aligned} \right\} \begin{aligned} k &= \frac{\sqrt{2mE}}{\hbar} \\ k' &= \frac{\sqrt{2m(E-V_o)}}{\hbar} \end{aligned} \quad (8.4.2)$$

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:

$$\begin{aligned} &\text{probability of finding between } x \text{ and } x + dx && (8.4.3) \\ : \left\{ \begin{aligned} \text{incident particle:} & \quad |\psi_{inc}(x)|^2 dx = |A|^2 dx \\ \text{reflected particle:} & \quad |\psi_{refl}(x)|^2 dx = |B|^2 dx \\ \text{transmitted particle:} & \quad |\psi_{trans}(x)|^2 dx = |C|^2 dx \end{aligned} \right. \end{aligned}$$

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:

$$T = \frac{|C|^2 k'}{|A|^2 k} \quad R = \frac{|B|^2 k}{|A|^2 k} = \frac{|B|^2}{|A|^2} \quad (8.4.4)$$

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:

$$T = \frac{4\sqrt{E(E-V_o)}}{(\sqrt{E} + \sqrt{E-V_o})^2} \quad R = \frac{(\sqrt{E} - \sqrt{E-V_o})^2}{(\sqrt{E} + \sqrt{E-V_o})^2} \quad (8.4.5)$$

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 :

$$T = \frac{4 \left(\frac{E}{V_o} \right) \left(\frac{E}{V_o} \pm 1 \right)}{\sin^2 \left(\sqrt{2m(E \pm V_o)} \frac{L}{\hbar} \right) + 4 \left(\frac{E}{V_o} \right) \left(\frac{E}{V_o} \pm 1 \right)} \quad R = \frac{\sin^2 \left(\sqrt{2m(E \pm V_o)} \frac{L}{\hbar} \right)}{\sin^2 \left(\sqrt{2m(E \pm V_o)} \frac{L}{\hbar} \right) + 4 \left(\frac{E}{V_o} \right) \left(\frac{E}{V_o} \pm 1 \right)} \quad (8.4.6)$$

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 :

$$\sqrt{2m(E \pm V_o)} \frac{L}{\hbar} = n\pi \Rightarrow E \pm V_o = \frac{n^2 \pi^2 \hbar^2}{2mL^2} = \frac{n^2 h^2}{8mL^2} \quad (8.4.7)$$

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:

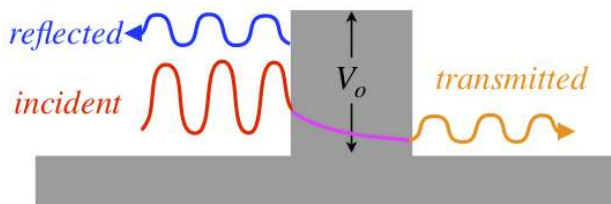
$$E \pm V_o = \frac{\hbar^2 k'^2}{2m} = \frac{4\hbar^2 \pi^2}{2m\lambda'^2} = \frac{n^2 \pi^2 \hbar^2}{2mL^2} \Rightarrow L = \frac{n\lambda'}{2} \quad (8.4.8)$$

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 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:

$$\sin^2 \left(\sqrt{2m(E - V_o)} \frac{L}{\hbar} \right) = \sin^2 \left(i \sqrt{2m(V_o - E)} \frac{L}{\hbar} \right) \quad (8.4.9)$$

Sine functions with imaginary arguments can be converted to hyperbolic sine functions multiplied by imaginary i . The proof is quick:

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \Rightarrow \sin ix = \frac{e^{-x} - e^x}{2i} = i \frac{e^x - e^{-x}}{2} = i \sinh x \quad (8.4.10)$$

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:

$$T = \frac{4 \left(\frac{E}{V_o} \right) \left(1 - \frac{E}{V_o} \right)}{\sinh^2 \left(\sqrt{2m(V_o - E)} \frac{L}{\hbar} \right) + 4 \left(\frac{E}{V_o} \right) \left(1 - \frac{E}{V_o} \right)} \quad R = \frac{\sinh^2 \left(\sqrt{2m(V_o - E)} \frac{L}{\hbar} \right)}{\sinh^2 \left(\sqrt{2m(V_o - E)} \frac{L}{\hbar} \right) + 4 \left(\frac{E}{V_o} \right) \left(1 - \frac{E}{V_o} \right)} \quad (8.4.11)$$

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.

This page titled 8.4: Tunneling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Tom Weideman.

8.5: 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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi_n(x) + \frac{1}{2} \kappa x^2 \psi_n(x) = E_n \psi_n(x) \quad (8.5.1)$$

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 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 must 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:

$$\psi(x) = Ae^{-\frac{1}{2}\alpha x^2} \quad (8.5.2)$$

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:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} [Ae^{-\frac{1}{2}\alpha x^2}] + \frac{1}{2}\kappa x^2 [Ae^{-\frac{1}{2}\alpha x^2}] = E_n [Ae^{-\frac{1}{2}\alpha x^2}] \quad (8.5.3)$$

Taking the derivatives and canceling the $e^{-\frac{1}{2}\alpha x^2}$ functions that appear in every term gives:

$$-\frac{\hbar^2}{2m} (-\alpha + \alpha^2 x^2) + \frac{1}{2}\kappa x^2 = E_n \quad (8.5.4)$$

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:

$$0 = -\frac{\hbar^2}{2m}\alpha^2 + \frac{1}{2}\kappa \Rightarrow \alpha = \frac{\sqrt{\kappa m}}{\hbar} \quad (8.5.5)$$

With the x^2 terms canceling, the constant terms are left behind, giving:

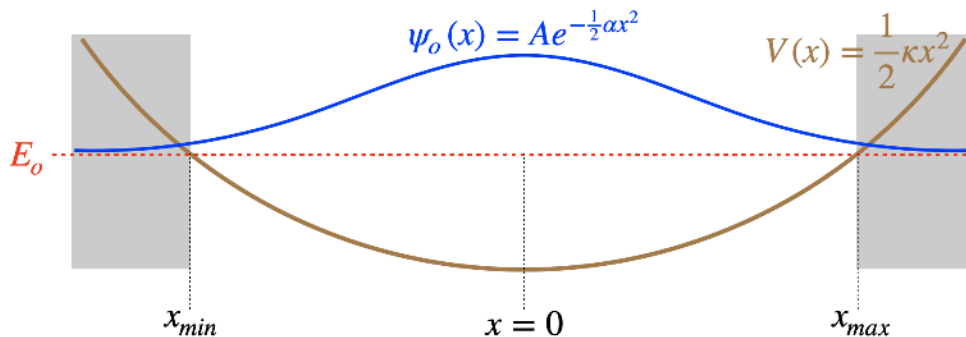
$$E_n = \frac{\hbar^2 \alpha}{2m} \quad (8.5.6)$$

Putting together these two results gives the energy in terms of given values:

$$E_n = \frac{\hbar^2 \frac{\sqrt{\kappa m}}{\hbar}}{2m} = \frac{1}{2} \hbar \sqrt{\frac{\kappa}{m}} \quad (8.5.7)$$

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 eigenstates have 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 8.5.1 – 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 (x_{min} and x_{max}) 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:

$$1 = \int_{-\infty}^{+\infty} |\psi(x)|^2 dx = A^2 \int_{-\infty}^{+\infty} e^{-\alpha x^2} dx = A^2 \left(\sqrt{\frac{\pi}{\alpha}} \right) \Rightarrow A = \left(\frac{\alpha}{\pi} \right)^{\frac{1}{4}} \quad (8.5.8)$$

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:

$$\begin{aligned} n = 0 : & \quad H_0(x) = Ax^0 \\ n = 1 : & \quad H_1(x) = Bx^1 \\ n = 2 : & \quad H_2(x) = Cx^0 + Dx^2 \\ n = 3 : & \quad H_3(x) = Ex^1 + Fx^3 \\ \vdots & \quad \vdots \end{aligned} \quad (8.5.9)$$

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 $H_n(x)$:

$$\psi_n(x) = \left(\frac{\beta}{2^n n! \sqrt{\pi}} \right)^{\frac{1}{2}} H_n(\beta x) e^{-\frac{(\beta x)^2}{2}} = \frac{1}{\sqrt{2^n n!}} H_n(\beta x) \psi_0(x) \quad (8.5.10)$$

For reasons of simplicity in some other constants, we have replaced the constant α with β^2 , so:

$$\beta \equiv \sqrt{\alpha} = \left(\frac{\kappa m}{\hbar^2} \right)^{\frac{1}{4}} \quad (8.5.11)$$

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:

$$\begin{aligned} H_0(\beta x) &= 1 \\ H_1(\beta x) &= 2\beta x \\ H_2(\beta x) &= 4\beta^2 x^2 - 2 \\ H_3(\beta x) &= 8\beta^3 x^3 - 12\beta x \end{aligned} \quad (8.5.12)$$

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:

$$E_n = \left(n + \frac{1}{2}\right) \hbar \omega_c, \quad n = 0, 1, 2, \dots, \quad (8.5.13)$$

$$\omega_c \equiv \sqrt{\frac{\kappa}{m}} = \text{angular frequency of classical oscillator}$$

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 $\hat{x}$. Given the reciprocal relationship we know exists between these two quantities (think of the Fourier transform and its inverse!), it is 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:

$$\begin{aligned} E_n = \langle E \rangle_n &= \langle KE + PE \rangle_n = \langle KE \rangle_n + \langle PE \rangle_n \Rightarrow \langle KE \rangle_n \\ &= \langle PE \rangle_n = \frac{1}{2} E_n = \frac{1}{2} \left(n + \frac{1}{2}\right) \hbar \omega_c \end{aligned} \quad (8.5.14)$$

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 these 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:

$$\begin{aligned} \langle x^2 \rangle_n &= \frac{2}{\kappa} \langle PE \rangle_n = \left(n + \frac{1}{2}\right) \hbar \left(\frac{\omega_c}{\kappa}\right) = \left(n + \frac{1}{2}\right) \hbar \frac{1}{\sqrt{\kappa m}} \\ \langle p^2 \rangle_n &= 2m \langle KE \rangle_n = \left(n + \frac{1}{2}\right) \hbar (m \omega_c) = \left(n + \frac{1}{2}\right) \hbar \sqrt{\kappa m} \end{aligned} \quad (8.5.15)$$

Plugging into the uncertainty equations:

$$\begin{aligned} \Delta x &= \sqrt{\langle x^2 \rangle} = \sqrt{\left(n + \frac{1}{2}\right) \hbar (\kappa m)^{-\frac{1}{2}}} \\ \Delta p &= \sqrt{\langle p^2 \rangle} = \sqrt{\left(n + \frac{1}{2}\right) \hbar (\kappa m)^{\frac{1}{2}}} \end{aligned} \quad (8.5.16)$$

Whenever we have the uncertainties for position and momentum, it is natural to want to test the uncertainty principle. So multiplying these together gives:

$$\Delta x \Delta p = \left(n + \frac{1}{2}\right) \hbar \quad (8.5.17)$$

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 9B, 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.

This page titled [8.5: The Quantum Harmonic Oscillator](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

8.6: 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):

$$\vec{F}_{electric} = \frac{kq_1q_2}{r^2} \hat{r} \quad (8.6.1)$$

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!):

$$k \equiv \frac{1}{4\pi\epsilon_o} \Rightarrow \vec{F}_{electric} = \frac{q_1q_2}{4\pi\epsilon_or^2} \hat{r}, \quad \epsilon_o = 8.85 \times 10^{-12} \frac{C^2}{Nm^2} \quad (8.6.2)$$

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:

$$V(r) = \frac{q_1q_2}{4\pi\epsilon_or} \quad (8.6.3)$$

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:

$$V(r) = -\frac{Ze^2}{4\pi\epsilon_or} \quad (8.6.4)$$

Now at last we can express the total energy of the electron as it orbits the nucleus:

$$E_{tot} = \frac{1}{2}mv^2 - \frac{Ze^2}{4\pi\epsilon_or} \quad (8.6.5)$$

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 encounter 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 electron 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:

$$m_e \frac{v^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2} \Rightarrow p^2 = (m_e v)^2 = \frac{m_e e^2}{4\pi\epsilon_0 r} \quad (8.6.6)$$

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:

$$\left. \begin{array}{l} \text{deBroglie : } p = \frac{h}{\lambda} \\ \text{Bohr : } 2\pi r = n\lambda \end{array} \right\} \Rightarrow p^2 = \left(\frac{nh}{2\pi r} \right)^2 = \frac{n^2 \hbar^2}{r^2} \quad (8.6.7)$$

Plugging this back into [Equation 8.6.6](#) and solving for the radius of the orbit gives:

$$r_n = n^2 \left(\frac{4\pi\epsilon_0 \hbar^2}{m_e e^2} \right) \equiv n^2 a_o \quad (8.6.8)$$

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 8.6.6 by twice the mass of the electron gives us its kinetic energy, and comparing the result with Equation 8.6.4 (with $Z = 1$ for the hydrogen atom) gives:

$$\begin{aligned}\frac{p^2}{2m} &= \frac{r}{2} \left(\frac{e^2}{4\pi\epsilon_0 r^2} \right) \Rightarrow KE = -\frac{1}{2}PE \Rightarrow E_{tot} \\ &= KE + PE = \frac{1}{2}PE = -\frac{e^2}{8\pi\epsilon_0 r}\end{aligned}\quad (8.6.9)$$

But the radii are quantized, so plugging this in gives quantized energy levels:

$$E_n = -\frac{e^2}{8\pi\epsilon_0 r_n} = -\frac{1}{n^2} \left(\frac{m_e e^4}{2(4\pi\epsilon_0)^2 \hbar^2} \right) \quad (8.6.10)$$

The energy spectrum is also quantized. The quantity in parentheses is a unit of energy known as a **Rydberg**, which $\approx 13.6\text{eV}$.

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:

$$hf = \Delta E = E_{n_1} - E_{n_2} = -13.6\text{eV} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \quad (8.6.11)$$

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.

This page titled [8.6: The Bohr Model of the Hydrogen Atom](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

CHAPTER OVERVIEW

9: Quantum Theory in Three Dimensions

[9.1: Schrödinger's Equation in 3-Dimensions](#)

[9.2: 3-Dimensional Models](#)

[9.3: Central Forces](#)

[9.4: Angular Momentum](#)

This page titled [9: Quantum Theory in Three Dimensions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

9.1: 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 $\vec{r}$.

$$\Psi(x, y, z, t) \rightarrow \Psi(\vec{r}, t) \quad (9.1.1)$$

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:

$$\begin{aligned} &\text{probability that particle is found in a tiny volume } dV \text{ located} \\ &\text{at position } \vec{r} = |\Psi(\vec{r}, t)|^2 dV \end{aligned} \quad (9.1.2)$$

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:

$$dV = \begin{array}{cc} \text{Cartesian} & \text{spherical} \\ dx \, dy \, dz & r^2 \, dr \, \sin \theta \, d\theta \, d\phi \end{array} \quad (9.1.3)$$

Naturally the normalization condition is:

$$\int_{\text{all space}} |\Psi(\vec{r}, t)|^2 dV = 1 \quad (9.1.4)$$

Written out explicitly for Cartesian and spherical coordinates, this is:

$$\begin{aligned} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz |\Psi(x, y, z, t)|^2 &= 1 \\ \int_0^{\infty} r^2 dr \int_0^{\pi} \sin \theta d\theta \int_0^{2\pi} d\phi |\Psi(r, \theta, \phi, t)|^2 &= 1 \end{aligned} \quad (9.1.5)$$

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:

$$\widehat{H} = \frac{1}{2m} (\hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2) + \widehat{V} = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V(x, y, z) \quad (9.1.6)$$

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):

$$\widehat{H} = -\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r}) \quad (9.1.7)$$

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:

$$\Psi(\vec{r}, t) = \psi(\vec{r}) e^{-i\omega t}, \quad \text{where:} \quad -\frac{\hbar^2}{2m} \nabla^2 \psi(\vec{r}) + V(\vec{r}) \psi(\vec{r}) = E \psi(\vec{r}), \quad \omega = \frac{E}{\hbar} \quad (9.1.8)$$

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:

$$\psi(x, y, z) = X(x) Y(y) Z(z) \quad (9.1.9)$$

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:

$$\begin{aligned} -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) X(x) Y(y) Z(z) + V(x, y, z) X(x) Y(y) Z(z) \\ = E X(x) Y(y) Z(z) \\ \Rightarrow \frac{1}{X} \frac{\partial^2 X}{\partial x^2} + \frac{1}{Y} \frac{\partial^2 Y}{\partial y^2} + \frac{1}{Z} \frac{\partial^2 Z}{\partial z^2} - \frac{2m}{\hbar^2} V(x, y, z) = -\frac{2mE}{\hbar^2} \end{aligned} \quad (9.1.10)$$

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:

$$f(x) + g(y) + h(z) = \text{constant} \quad (9.1.11)$$

We can plug any x value we like into $f(x)$, without changing any of the y or z values. For this equation to remain correct, it must mean that $f(x)$ is the same constant value for all choices of x . The same argument can be made for $g(y)$ and $h(z)$. 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:

$$\frac{1}{X} \frac{d^2 X}{dx^2} = -k_x^2, \quad \frac{1}{Y} \frac{d^2 Y}{dy^2} = -k_y^2, \quad \frac{1}{Z} \frac{d^2 Z}{dz^2} = -k_z^2, \quad k_x^2 + k_y^2 + k_z^2 = \frac{2mE}{\hbar^2} \quad (9.1.12)$$

The plane wave solutions of each of the separate differential equations are:

$$X(x) = A_x e^{ik_x x}, \quad Y(y) = A_y e^{ik_y y}, \quad Z(z) = A_z e^{ik_z z} \quad (9.1.13)$$

Reconstructing the full momentum eigenstate wave function, we get:

$$\psi(x, y, z) = X(x) Y(y) Z(z) = A e^{i(k_x x + k_y y + k_z z)} \quad (9.1.14)$$

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:

$$\psi(\vec{r}) = A e^{i\vec{k} \cdot \vec{r}}, \quad \vec{k} = \frac{\vec{p}}{\hbar}, \quad E = \frac{p^2}{2m} = \frac{\hbar^2}{2m} (\vec{k} \cdot \vec{k}) \quad (9.1.15)$$

This plane wave is also an energy eigenstate, so the full time-dependent wave function can also be written:

$$\psi(\vec{r}, t) = A e^{i(\vec{k} \cdot \vec{r} - \omega t)}, \quad \omega = \frac{E}{\hbar} \quad (9.1.16)$$

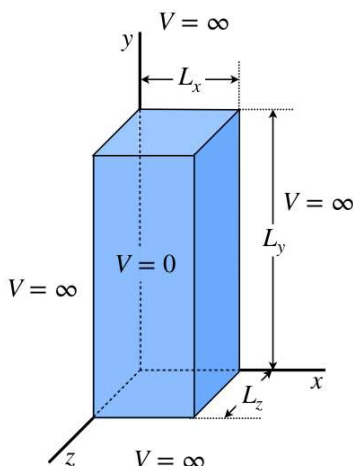
This page titled [9.1: Schrödinger's Equation in 3-Dimensions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

9.2: 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:

$$V(x, y, z) = \begin{cases} 0 & 0 < x < L_x \text{ and } 0 < y < L_y \text{ and } 0 < z < L_z \\ \infty & \text{elsewhere} \end{cases} \quad (9.2.1)$$

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:

$$X(x) = A_x \sin \frac{n_x \pi x}{L_x} \quad Y(y) = A_y \sin \frac{n_y \pi y}{L_y} \quad Z(z) = A_z \sin \frac{n_z \pi z}{L_z}, \quad (9.2.2)$$

$n_x, n_y, n_z = 1, 2, \dots$

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.

$$\psi_{n_x n_y n_z}(x, y, z) = \sqrt{\frac{8}{L_x L_y L_z}} \sin \frac{n_x \pi x}{L_x} \sin \frac{n_y \pi y}{L_y} \sin \frac{n_z \pi z}{L_z}, \quad n_x, n_y, n_z = 1, 2, \dots \quad (9.2.3)$$

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:

$$E_{n_x n_y n_z} = \left(\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2} \right) \frac{h^2}{8m}, \quad n_x, n_y, n_z = 1, 2, \dots \quad (9.2.4)$$

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:

$$V(x, y, z) = \frac{1}{2}\kappa_x x^2 + \frac{1}{2}\kappa_y y^2 + \frac{1}{2}\kappa_z z^2 \quad (9.2.5)$$

This sort of model might be useful for crystal lattices where the bonds differ along the different dimensions. Plugging this into [Equation 9.1.10](#) results in an equation with three separated functions again:

$$\left[\frac{1}{X} \frac{\partial^2 X}{\partial x^2} + \frac{\kappa_x m}{\hbar^2} x^2 \right] + \left[\frac{1}{Y} \frac{\partial^2 Y}{\partial y^2} + \frac{\kappa_y m}{\hbar^2} y^2 \right] + \left[\frac{1}{Z} \frac{\partial^2 Z}{\partial z^2} + \frac{\kappa_z m}{\hbar^2} z^2 \right] = -\frac{2mE}{\hbar^2} \quad (9.2.6)$$

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:

$$E_{n_x n_y n_z} = \left(n_x + \frac{1}{2} \right) \hbar \sqrt{\frac{\kappa_x}{m}} + \left(n_y + \frac{1}{2} \right) \hbar \sqrt{\frac{\kappa_y}{m}} + \left(n_z + \frac{1}{2} \right) \hbar \sqrt{\frac{\kappa_z}{m}} \quad (9.2.7)$$

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:

$$E_{n_x n_y n_z} = \left(n_x + n_y + n_z + \frac{3}{2} \right) \hbar \omega_c, \quad \omega_c = \sqrt{\frac{\kappa}{m}} \quad (9.2.8)$$

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.

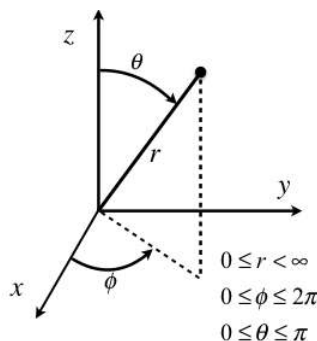
This page titled [9.2: 3-Dimensional Models](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

9.3: 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 9.1.7) in spherical coordinates. This we can simply look up (deriving it from the transformation between coordinate systems is no fun):

$$\nabla^2 = \frac{1}{r^2} \left[\frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \left(\frac{1}{\sin \theta} \right) \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \left(\frac{1}{\sin^2 \theta} \right) \frac{\partial^2}{\partial \phi^2} \right] \quad (9.3.1)$$

The stationary-state Schrodinger equation in spherical coordinates is therefore:

$$-\frac{\hbar^2}{2mr^2} \left[\frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \left(\frac{1}{\sin \theta} \right) \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \left(\frac{1}{\sin^2 \theta} \right) \frac{\partial^2}{\partial \phi^2} \right] \psi(r, \theta, \phi) + V(r, \theta, \phi) \psi(r, \theta, \phi) = E \psi(r, \theta, \phi) \quad (9.3.2)$$

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:

$$\begin{aligned} & \left(\frac{1}{\sin \theta} \right) \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) \psi(r, \theta, \phi) + \left(\frac{1}{\sin^2 \theta} \right) \frac{\partial^2}{\partial \phi^2} \psi(r, \theta, \phi) \\ &= \frac{2mr^2}{\hbar^2} [V(r, \theta, \phi) - E] \psi(r, \theta, \phi) - \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) \psi(r, \theta, \phi) \end{aligned} \quad (9.3.3)$$

Now follow our two-step separation trick: Write the wave function as a product of single-variable functions...

$$\psi(r, \theta, \phi) = R(r) \Theta(\theta) \Phi(\phi) \quad (9.3.4)$$

...and divide the whole equation by the wave function:

$$\begin{aligned} & \frac{1}{\Theta} \left(\frac{1}{\sin \theta} \right) \frac{d}{d\theta} \left(\sin \theta \frac{d}{d\theta} \right) \Theta + \frac{1}{\Phi} \left(\frac{1}{\sin^2 \theta} \right) \frac{d^2}{d\phi^2} \Phi \\ &= \frac{2mr^2}{\hbar^2} [V(r, \theta, \phi) - E] - \frac{1}{R} \frac{d}{dr} \left(r^2 \frac{d}{dr} \right) R \end{aligned} \quad (9.3.5)$$

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**:

$$-\frac{\hbar^2}{2mr^2} \frac{d}{dr} \left(r^2 \frac{d}{dr} \right) R(r) + \frac{\hbar^2 C_1}{2mr^2} R(r) + V(r) R(r) = E R(r) \quad (9.3.6)$$

Plugging C_1 into the right-hand side of Equation 9.3.5, we now set out to separate the angular functions:

$$\frac{1}{\Theta} \left(\frac{1}{\sin \theta} \right) \frac{d}{d\theta} \left(\sin \theta \frac{d}{d\theta} \right) \Theta + \frac{1}{\Phi} \left(\frac{1}{\sin^2 \theta} \right) \frac{d^2}{d\phi^2} \Phi = C_1 \quad (9.3.7)$$

Multiply the equation by $\sin^2 \theta$ and collect the functions of each variable to get:

$$\frac{1}{\Theta} \left[\sin \theta \frac{d}{d\theta} \left(\sin \theta \frac{d}{d\theta} \right) \Theta - C_1 \sin^2 \theta \right] = -\frac{1}{\Phi} \frac{d^2}{d\phi^2} \Phi \quad (9.3.8)$$

Separated at last, we can state that both sides of this equation equal a constant (which we will call " C_2 "), giving us two more differential equations in single variables, the *polar equation* (in θ) and the *azimuthal equation* (in ϕ):

$$\sin \theta \frac{d}{d\theta} \left(\sin \theta \frac{d}{d\theta} \right) \Theta (\theta) - (C_1 \sin^2 \theta + C_2) \Theta (\theta) = 0 \quad (9.3.9)$$

$$\frac{d^2}{d\phi^2} \Phi (\phi) + C_2 \Phi (\phi) = 0 \quad (9.3.10)$$

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:

$$\Phi (\phi) = A e^{\pm i \sqrt{C_2} \phi} \quad (9.3.11)$$

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:

$$\Phi_{m_l} (\phi) = A e^{i m_l \phi}, \quad m_l = 0, \pm 1, \pm 2, \dots \quad (9.3.12)$$

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:

$$C_1 = -l(l+1), \quad l = 0, 1, 2, \dots \quad (9.3.13)$$

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):

$$Y_{lm} (\theta, \phi) = \Theta_{lm_l} (\theta) \Phi_{m_l} (\phi) \quad (9.3.14)$$

[Note: In the literature, the indices defining spherical harmonics are frequently written with the m_l index raised: $Y_l^{m_l}$.]

For a given value of l , it so happens that m_l can take on only a limited number of values:

$$m_l = 0, \pm 1, \pm 2, \dots, \pm l \quad (9.3.15)$$

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:

$$-\frac{\hbar^2}{2mr^2} \frac{d}{dr} \left(r^2 \frac{d}{dr} \right) R(r) - \frac{\hbar^2 l(l+1)}{2mr^2} R(r) + V(r) R(r) = E R(r) \quad (9.3.16)$$

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:

$$\psi_{nlm_l}(r, \theta, \phi) = AR_{nl}(r)Y_{lm_l}(\theta, \phi) \quad (9.3.17)$$

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:

$$\psi(r) = A \frac{e^{\pm ikr}}{r}, \quad E = \frac{\hbar^2 k^2}{2m} \quad (9.3.18)$$

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 $e^{-i\omega t}$.

This page titled [9.3: Central Forces](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Tom Weideman](#).

9.4: 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 its 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:

$$\left. \begin{array}{l} x \rightarrow -x \\ y \rightarrow -y \\ z \rightarrow -z \end{array} \right\} \rightarrow \vec{L} = \vec{r} \times \vec{p} = (-\vec{r}) \times (-\vec{p}) = +\vec{L} \quad (9.4.1)$$

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:

$$\widehat{KE} \equiv \frac{1}{2m} (\hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2) = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \quad (9.4.2)$$

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:

$$\begin{aligned} \vec{L} = \vec{r} \times \vec{p} &= (x \hat{i} + y \hat{j} + z \hat{k}) \times (p_x \hat{i} + p_y \hat{j} + p_z \hat{k}) \\ &= (yp_z - zp_y) \hat{i} + (zp_x - xp_z) \hat{j} + (xp_y - yp_x) \hat{z} \end{aligned} \quad (9.4.3)$$

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:

$$\widehat{L}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x = -i\hbar \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right) \quad (9.4.4)$$

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:

$$\widehat{L}_x \widehat{L}_y \Psi = \left[-i\hbar \left(y \frac{\partial}{\partial z} - z \frac{\partial}{\partial y} \right) \right] \left[-i\hbar \left(z \frac{\partial}{\partial x} - x \frac{\partial}{\partial z} \right) \right] \Psi$$

$$\begin{aligned}
 &= -\hbar^2 \left[y \frac{\partial}{\partial x} \Psi + yz \frac{\partial^2}{\partial x \partial z} \Psi + xz \frac{\partial^2}{\partial y \partial z} \Psi - z^2 \frac{\partial^2}{\partial x \partial y} \Psi - xy \frac{\partial^2}{\partial z^2} \Psi \right] \\
 \widehat{L}_y \widehat{L}_x \Psi &= \left[-i\hbar \left(z \frac{\partial}{\partial x} - x \frac{\partial}{\partial z} \right) \right] \left[-i\hbar \left(y \frac{\partial}{\partial z} - z \frac{\partial}{\partial y} \right) \right] \Psi \\
 &= -\hbar^2 \left[x \frac{\partial}{\partial y} \Psi + yz \frac{\partial^2}{\partial x \partial z} \Psi + xz \frac{\partial^2}{\partial y \partial z} \Psi - z^2 \frac{\partial^2}{\partial x \partial y} \Psi - xy \frac{\partial^2}{\partial z^2} \Psi \right]
 \end{aligned} \tag{9.4.5}$$

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:

$$(\widehat{L}_x \widehat{L}_y - \widehat{L}_y \widehat{L}_x) \Psi = \hbar^2 \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right) \Psi \tag{9.4.6}$$

The right side of this equation looks familiar – it is very nearly the operator of the z -component of angular momentum ([Equation 9.4.4](#)). Indeed we have:

$$\widehat{L}_x \widehat{L}_y - \widehat{L}_y \widehat{L}_x = i\hbar \widehat{L}_z \tag{9.4.7}$$

There is a shorthand notation frequently used in such situations, called the **commutator** of the operators $\widehat{A}$ and $\widehat{B}$:

$$[\widehat{A}, \widehat{B}] \equiv \widehat{A}\widehat{B} - \widehat{B}\widehat{A} \tag{9.4.8}$$

Thanks to the cyclical nature of the cross-product, we can similarly construct all the components from the other two, so we have:

$$[\widehat{L}_x, \widehat{L}_y] = i\hbar \widehat{L}_z, \quad [\widehat{L}_y, \widehat{L}_z] = i\hbar \widehat{L}_x, \quad [\widehat{L}_z, \widehat{L}_x] = i\hbar \widehat{L}_y \tag{9.4.9}$$

These are known as the **commutation relations** of the angular momentum component operators.

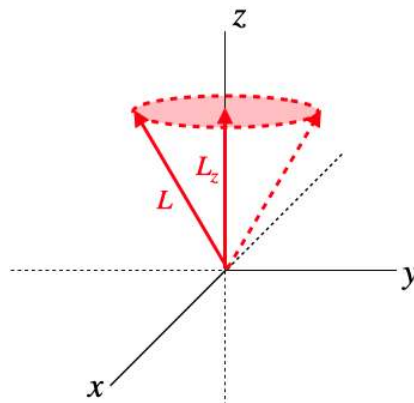
Magnitude of Angular Momentum

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:

$$[\widehat{L}^2, \widehat{L}_x] = [\widehat{L}^2, \widehat{L}_y] = [\widehat{L}^2, \widehat{L}_z] = 0 \quad \widehat{L}^2 = \widehat{L}_x^2 + \widehat{L}_y^2 + \widehat{L}_z^2 \tag{9.4.10}$$

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 9.4.1 – 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 $\hat{L}_z$ 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:

$$\hat{L}_z = -i\hbar \frac{\partial}{\partial \phi} \quad (9.4.11)$$

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:

$$\hat{L}_z \Psi_{nlm_l}(r, \theta, \phi) = -i\hbar \frac{\partial}{\partial \phi} [R_{nl}(r) \Theta_{lm_l}(\theta) e^{im_l \phi}] = m_l \hbar \Psi_{nlm_l}(r, \theta, \phi) \quad (9.4.12)$$

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 $\hat{L}^2$ is considerably more complicated than the one for $\hat{L}_z$, 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:

$$\hat{L}^2 \Psi_{nlm_l}(r, \theta, \phi) = \hat{L}^2 [R_{nl}(r) \Theta_{lm_l}(\theta) e^{im_l \phi}] = l(l+1) \hbar^2 \Psi_{nlm_l}(r, \theta, \phi) \quad (9.4.13)$$

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 $|\vec{L}| = \sqrt{l(l+1)}\hbar$.

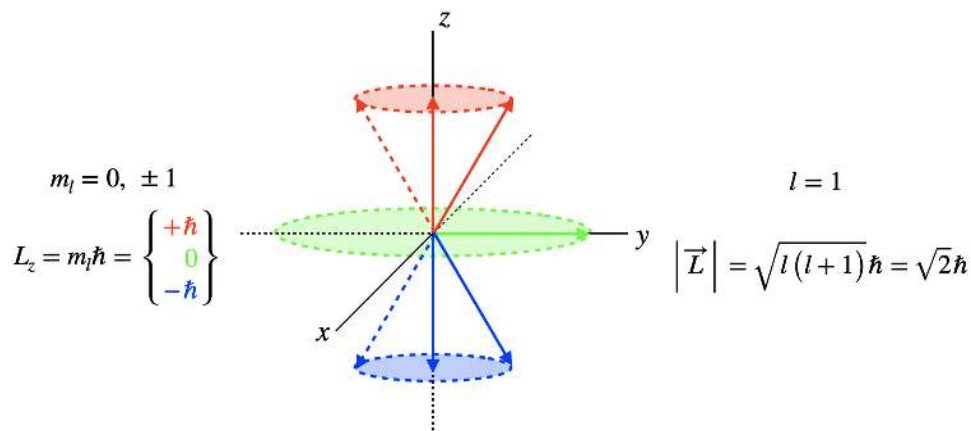
To see further justification for this eigenvalue for the square of the total angular momentum, take a look at the radial Schrödinger equation (Equation 9.3.16). The terms on the left side of the equation that do not include the potential energy must be the kinetic energy (since the right hand side is 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 9.3.8. 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 9.3.15, 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:

$$\text{number of quantized angular momentum states} = 2l + 1 \quad (9.4.14)$$

For example, for $l = 1$:

Figure 9.4.2 – 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 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 , and m_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 9.3.15), 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:

$$l = 0, 1, 2, \dots, n - 1 \quad (9.4.15)$$

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:

$$\begin{aligned} \text{degeneracy of } n^{\text{th}} \text{ energy eigenstate} &= [1 \ (l = 0, \ m = 0)] \\ &+ [3 \ (l = 1, \ m = 0, \pm 1)] + \dots + [2l + 1 \ (l = n - 1, \ m = 0, \pm 1, \dots, \pm l)] = n^2 \end{aligned} \quad (9.4.16)$$

This page titled 9.4: Angular Momentum is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Tom Weideman.

- 1.7: Examples of 2-Dimensional Motion by Tom Weideman is licensed CC BY-SA 4.0. Original source: [native](#).

Index

A

accidental degeneracies
[9.2: 3-Dimensional Models](#)

D

degeneracies
[9.2: 3-Dimensional Models](#)

G

Gedankenexperiment
[2.2: The Nature of Time](#)
gyromagnetic ratio
[10.2: It's Not Rotating!](#)

I

isotropic harmonic oscillator
[9.2: 3-Dimensional Models](#)

P

proper time
[2.2: The Nature of Time](#)

T

time dilation
[2.2: The Nature of Time](#)

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: UCD: Physics 9D – Modern Physics

Webpages: 61

All licenses found:

- [CC BY-SA 4.0](#): 83.6% (51 pages)
- [Undeclared](#): 16.4% (10 pages)

By Page

- UCD: Physics 9D – Modern Physics - [CC BY-SA 4.0](#)
 - Front Matter - [Undeclared](#)
 - [TitlePage](#) - [Undeclared](#)
 - [InfoPage](#) - [Undeclared](#)
 - [Table of Contents](#) - [Undeclared](#)
 - [Licensing](#) - [Undeclared](#)
 - 1: Sound - [CC BY-SA 4.0](#)
 - [1.1: Fundamentals of Sound](#) - [CC BY-SA 4.0](#)
 - [1.2: Doppler Effect](#) - [CC BY-SA 4.0](#)
 - [1.3: Interference Effects](#) - [CC BY-SA 4.0](#)
 - 2: Foundations of Special Relativity - [Undeclared](#)
 - [2.1: The Relativity Principle](#) - [CC BY-SA 4.0](#)
 - [2.2: The Nature of Time](#) - [CC BY-SA 4.0](#)
 - [2.3: More Thought Experiments](#) - [CC BY-SA 4.0](#)
 - [2.4: Paradoxes](#) - [CC BY-SA 4.0](#)
 - 3: Kinematics in Special Relativity - [CC BY-SA 4.0](#)
 - [3.1: Spacetime Diagrams](#) - [CC BY-SA 4.0](#)
 - [3.2: Lorentz Transformation](#) - [CC BY-SA 4.0](#)
 - [3.3: Velocity Addition](#) - [CC BY-SA 4.0](#)
 - [3.4: Electricity and Magnetism](#) - [CC BY-SA 4.0](#)
 - 4: Dynamics in Special Relativity - [CC BY-SA 4.0](#)
 - [4.1: Momentum Conservation](#) - [CC BY-SA 4.0](#)
 - [4.2: Energy Conservation](#) - [CC BY-SA 4.0](#)
 - 5: "Wait, what?!" Experiments Reveal Cracks in Our Understanding - [CC BY-SA 4.0](#)
 - [5.1: Blackbody Radiation](#) - [CC BY-SA 4.0](#)
 - [5.2: The Photoelectric Effect](#) - [CC BY-SA 4.0](#)
 - [5.3: Compton Effect](#) - [CC BY-SA 4.0](#)
 - [5.4: Double-Slit Experiment](#) - [CC BY-SA 4.0](#)
 - [5.5: Matter Has Wave Properties, Too!](#) - [CC BY-SA 4.0](#)
 - [5.6: Summarizing this Wave/Particle Mess](#) - [CC BY-SA 4.0](#)
 - 6: The Universe is Inherently Probabilistic - [CC BY-SA 4.0](#)
 - [6.1: Basics of Probability Theory](#) - [CC BY-SA 4.0](#)
 - [6.2: Continuous Probability Distributions and Probability Density](#) - [CC BY-SA 4.0](#)
 - [6.3: The Uncertainty of Random Outcomes](#) - [CC BY-SA 4.0](#)
 - [6.4: Physical Measurements with Random Outcomes](#) - [CC BY-SA 4.0](#)
 - [6.5: Incompatible Measurements](#) - [CC BY-SA 4.0](#)
 - 7: Matter Waves - [CC BY-SA 4.0](#)
 - [7.1: The Schrödinger Wave Equation](#) - [CC BY-SA 4.0](#)
 - [7.2: States of Definite Energy](#) - [CC BY-SA 4.0](#)
 - [7.3: Operators and Observables](#) - [CC BY-SA 4.0](#)
 - [7.4: Eigenstates and Eigenvalues](#) - [CC BY-SA 4.0](#)
 - 8: One-Dimensional Models - [CC BY-SA 4.0](#)
 - [8.1: Particle-in-a-Box, Part 1](#) - [CC BY-SA 4.0](#)
 - [8.2: Particle-in-a-Box, Part 2](#) - [CC BY-SA 4.0](#)
 - [8.3: The Finite Square Well](#) - [CC BY-SA 4.0](#)
 - [8.4: Tunneling](#) - [CC BY-SA 4.0](#)
 - [8.5: The Quantum Harmonic Oscillator](#) - [CC BY-SA 4.0](#)
 - [8.6: The Bohr Model of the Hydrogen Atom](#) - [CC BY-SA 4.0](#)
 - 9: Quantum Theory in Three Dimensions - [CC BY-SA 4.0](#)
 - [9.1: Schrödinger's Equation in 3-Dimensions](#) - [CC BY-SA 4.0](#)
 - [9.2: 3-Dimensional Models](#) - [CC BY-SA 4.0](#)
 - [9.3: Central Forces](#) - [CC BY-SA 4.0](#)
 - [9.4: Angular Momentum](#) - [CC BY-SA 4.0](#)
 - 10: Intrinsic Angular Momentum – "Spin" - [CC BY-SA 4.0](#)
 - [10.1: Measuring Angular Momentum](#) - [CC BY-SA 4.0](#)
 - [10.2: It's Not Rotating!](#) - [CC BY-SA 4.0](#)
 - [10.3: Fermions and Bosons](#) - [CC BY-SA 4.0](#)
 - Back Matter - [Undeclared](#)
 - [Index](#) - [Undeclared](#)
 - [Glossary](#) - [Undeclared](#)
 - [Detailed Licensing](#) - [Undeclared](#)