UCD: PHYSICS 9D - MODERN PHYSICS

Tom Weideman
University of California, Davis

LibreTexts™

University of California, Davis

UCD: Physics 9D - Modern Physics

Tom Weideman

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

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

TABLE OF CONTENTS

Licensing

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: Light as a Particle

- 5.1: Blackbody Radiation
- 5.2: The Photoelectric Effect
- 5.3: Compton Effect
- 5.4: Double-Slit Experiment

6: Matter as a Wave

- 6.1: From Light to Electrons
- 6.2: Interpreting Matter Waves

7: Quantum Mechanics in 1-Dimension

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

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 19^{th} and 20^{th} 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).

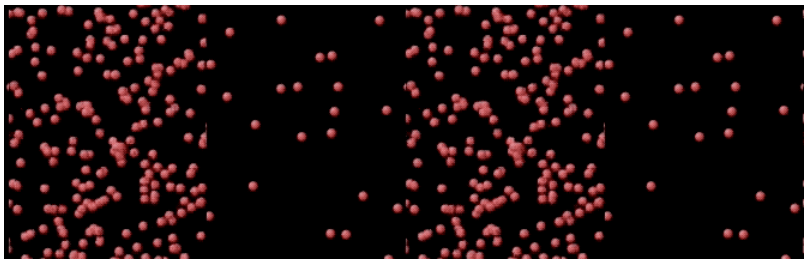
Sound Waves in Air

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 $343 m/s$. 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:

$$v_{sound\ in\ fluid} = \sqrt{\frac{B}{\rho}} \tag{1.1.1}$$

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:

$$v_{sound\ in\ solid} = \sqrt{\frac{Y}{\rho}} \tag{1.1.2}$$

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:

$$I = \frac{A^2}{2\sqrt{\rho B}} \tag{1.1.3}$$

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 $1 W/m^2$. 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):

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

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!):

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

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 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:

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

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

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

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:

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

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:

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

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):*

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

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

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

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:

$$\lambda = \frac{v}{f_s} \tag{1.2.4}$$

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:

$$T_r = \frac{\lambda}{v + v_r} \tag{1.2.5}$$

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

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

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:

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

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 3/4 of the source frequency:

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

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):

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

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:

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

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:	$109kHz$

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

$$\left.\begin{aligned} f_{echo} &= \left(\frac{v}{v - v_{drone}}\right) f_{emitted} \qquad [\text{moving drone sends sound}] \\ f_{received} &= \left(\frac{v + v_{drone}}{v}\right) f_{echo} \qquad [\text{moving drone receives echoed sound}] \end{aligned}\right\} \Rightarrow \quad f_{received} = \left(\frac{v + v_{drone}}{v - v_{drone}}\right) f_{emitted}$$

$$= \left(\frac{344\frac{m}{s} + 24\frac{m}{s}}{344\frac{m}{s} - 24\frac{m}{s}}\right)(100kHz) = 115Hz$$

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):

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

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:

$$v_{ufo} = -0.0268\ v = -9.2\frac{m}{s} \tag{1.2.9}$$

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).

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 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?

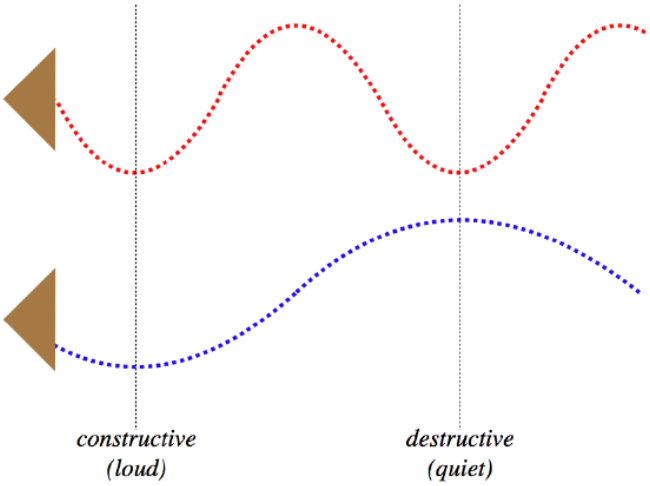
Solution

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:

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

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:

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

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:

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

Now apply a trigonometric identity:

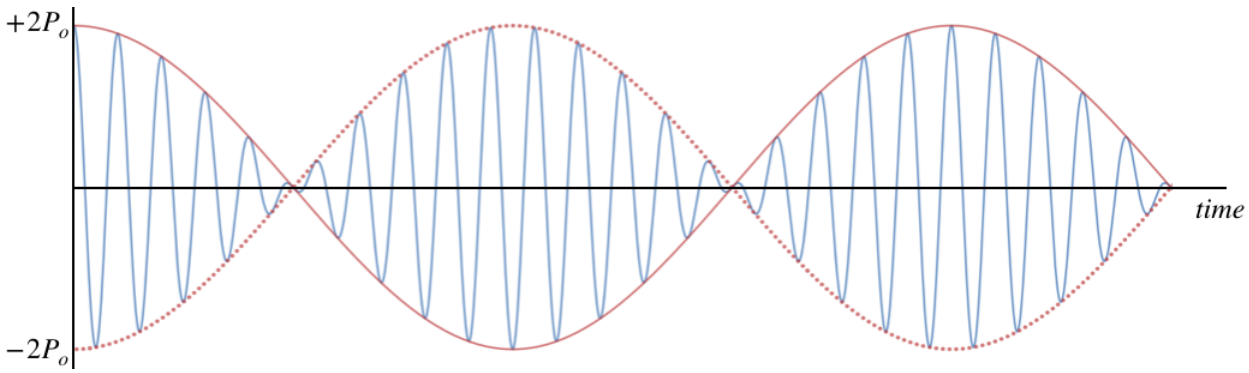
$$\cos X + \cos Y = 2\cos\left(\frac{X-Y}{2}\right)\cos\left(\frac{X+Y}{2}\right) \tag{1.3.4}$$

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

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:

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

[*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:

$$f_{carrier} = \frac{f_1 + f_2}{2} \tag{1.3.6}$$

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 $4.0\frac{cm}{s}$. 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:

$$\text{distance between nodes} = v_{bug}t = \frac{\lambda}{2} \quad \Rightarrow \quad \lambda = 2\left(4.0\frac{cm}{s}\right)(2.0s) = 16cm \quad \Rightarrow \quad f = \frac{v}{\lambda} = \frac{344\frac{m}{s}}{0.16m} = 2150Hz$$

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):*

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

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:

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

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:

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

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

2: Foundations of Special Relativity

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

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)

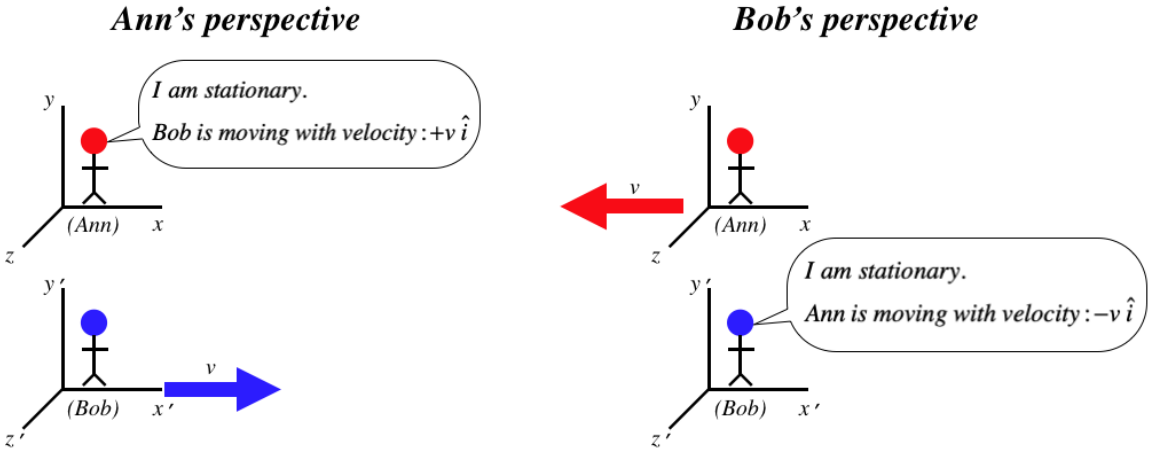
We have a simple experiment for testing whether our frame is inertial, but it doesn't tell us whether our frame is stationary or moving in a straight line at a constant speed, because when we release the pencil under these circumstances, it remains stationary from our perspective in both cases. So what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

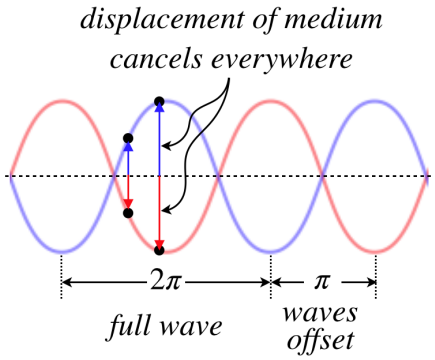
Waves are phenomena which transfer energy through space by the means of a self-propagating disturbance of a medium (the stuff that is "waving"). A wave on a string carries energy along because a piece of the string displaces, which pulls on an adjacent piece of string, displacing it, and so on. Surface 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.

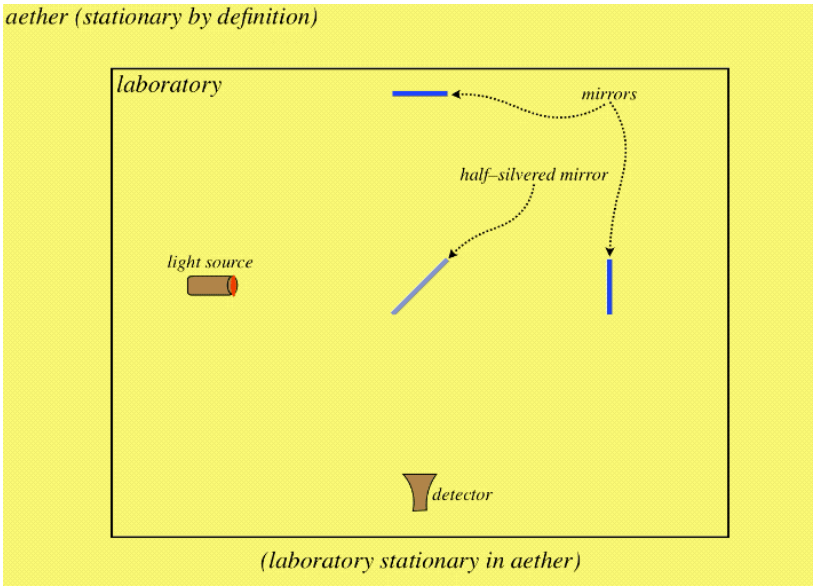
But far more puzzling than this was the fact that no one could seem to come up with any experimental evidence of the existence of a medium for light (which they referred to as the *luminiferous aether*). The most convincing null result of such a search was performed by two American physicists Albert 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

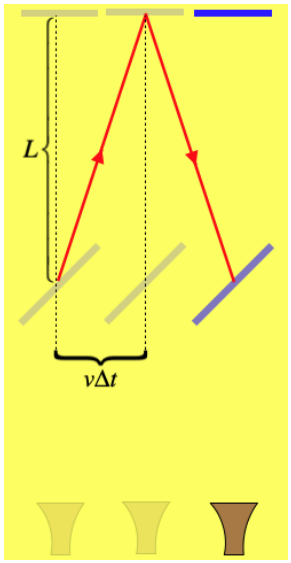


If this device is moving through aether (let's say in a direction parallel to one of these directions), then one of the arms of the interferometer moves parallel to the direction of the light during the journey, while the other moves perpendicular to it. These two journeys are not equal in distance,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



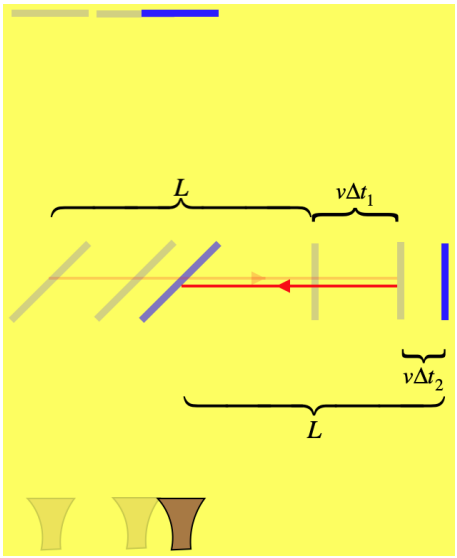
The transverse beam (the one perpendicular to the laboratory's motion) follows a diagonal path in both directions, the distance of which can be determined from the Pythagorean Theorem:

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

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:

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

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:

$$\left.\begin{aligned} s_1 = L + v\Delta t_1 = c\Delta t_1 \quad &\Rightarrow \quad \Delta t_1 = \frac{L}{c-v} \\ s_2 = L - v\Delta t_2 = c\Delta t_2 \quad &\Rightarrow \quad \Delta t_2 = \frac{L}{c+v} \end{aligned}\right\} \quad \Delta t_1 + \Delta t_2 = \frac{2LC}{c^2-v^2} \quad \Rightarrow \quad s = c\left(\Delta t_1 + \Delta t_2\right) = \frac{2L}{1-\frac{v^2}{c^2}} \tag{2.1.3}$$

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:

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

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.

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*.

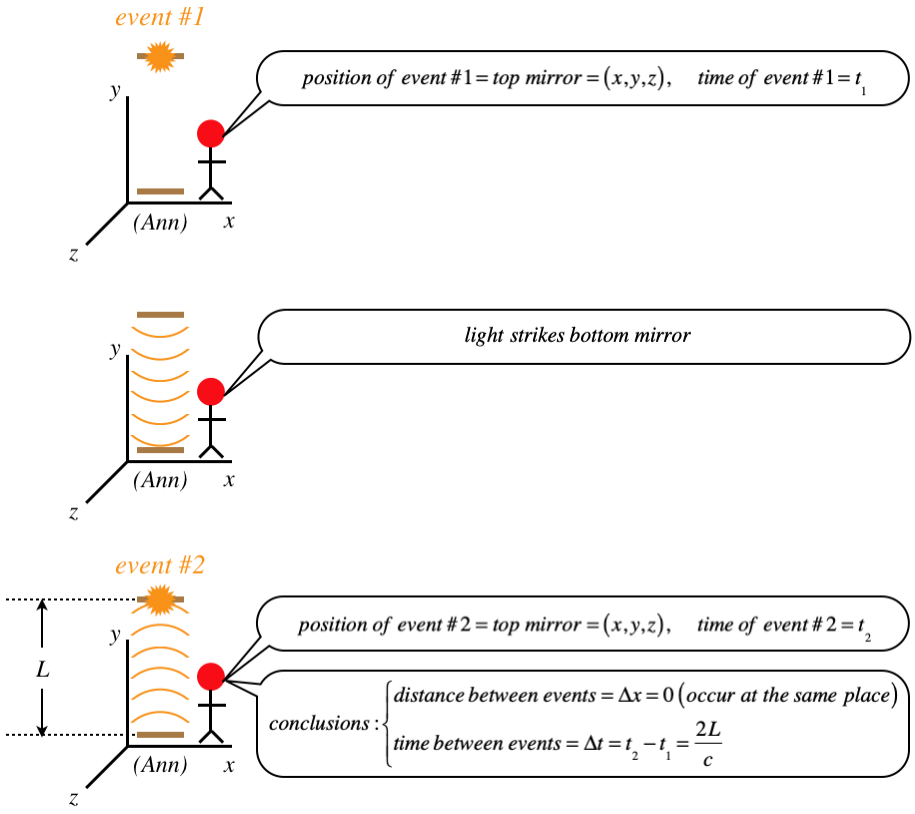
In the context of special relativity, a spacetime event is an instantaneous occurrence at a specific point in space and at a specific moment in time. A single point on a stationary light bulb as it dims defines a specified location, but it is not an event because the dimming process does not occur at 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

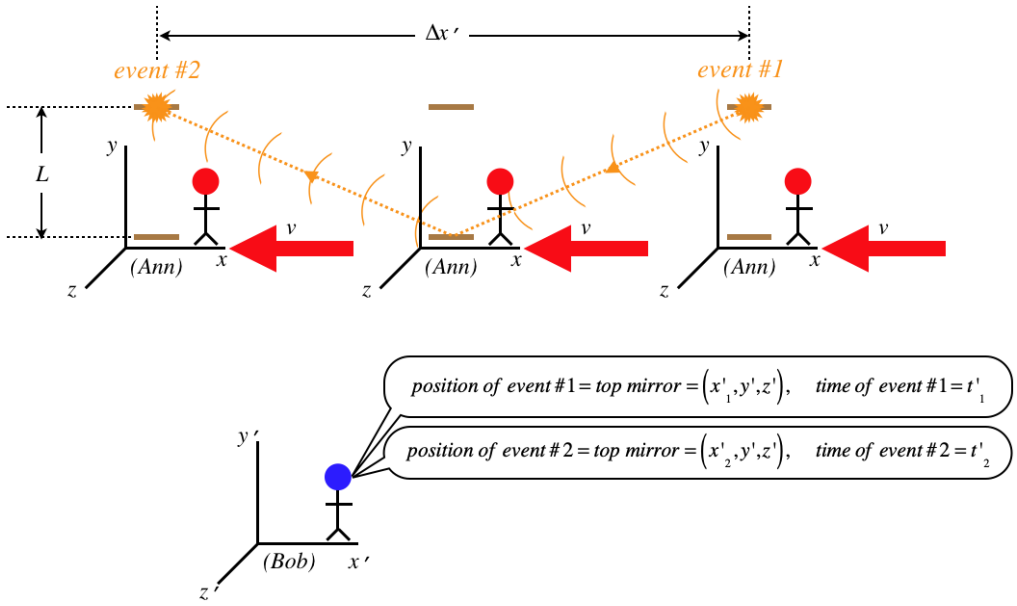


Okay, so we have used the two events to determine the time span between them according to Ann. The goal of relativity is to describe what a second observer measures for a physical process given what the first observer measured. So now we introduce Bob,

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:

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

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

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

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 make 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 it 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:

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

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 $\Delta t = t_2 - t_1$:

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

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

$$\Delta s^2 \equiv c^2\Delta t^2 - \left(\Delta x^2 + \Delta y^2 + \Delta z^2\right) \tag{2.2.4}$$

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:

$$\Delta s = c\Delta\tau \tag{2.2.5}$$

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:

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

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.

$$\Delta s^2_{Ann} = c^2\Delta t^2_{Ann} - \left(\Delta x^2_{Ann} + \cancelto{0}{\Delta y^2_{Ann}} + \cancelto{0}{\Delta z^2_{Ann}}\right)$$

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

$$\Delta s^2_{Bob} = c^2\Delta t^2_{Bob} - \left(\Delta x^2_{Bob} + \cancelto{0}{\Delta y^2_{Bob}} + \cancelto{0}{\Delta z^2_{Bob}}\right)$$

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:

$$\Delta x_{Bob} = v\Delta t_{Bob} \quad \Rightarrow \quad \Delta s^2_{Bob} = c^2\Delta t^2_{Bob} - \left(v\Delta t_{Bob}\right)^2 = \left(c^2 - v^2\right)\Delta t^2_{Bob}$$

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

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

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

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 \(\lambda\. 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:

$$f = \frac{1}{T} \tag{2.3.1}$$

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):

$$c = \frac{\lambda}{T} = \lambda f \tag{2.3.2}$$

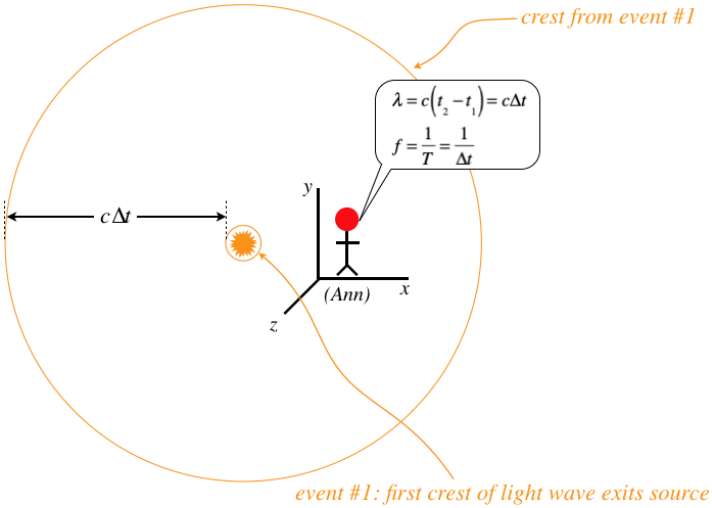
In the analysis to come, we will represent the "crests" of light waves with circles, so that the distance between these circles is the wavelength. We'll start with the basic phenomenon of doppler effect. The two gifs that follow apply equally to light or sound. The main idea is to note that when there 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 1.3.1 – No Relative Motion of Source and Receiver

Figure 1.3.2 – Approaching Relative Motion (Receiver's Perspective)

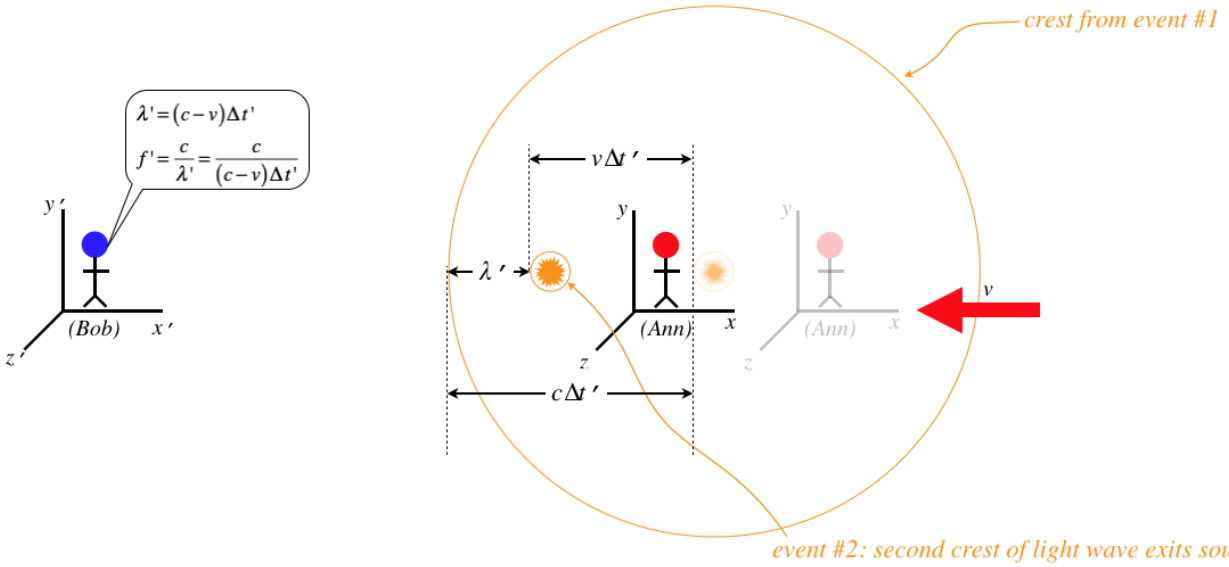
Now let's put Ann and Bob into the roles of source and receiver, respectively. We start with what Ann measures for the light source that remains stationary in her frame:

Figure 1.3.3 – 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 1.3.4 – 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 interial 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:

$$f' = \frac{c}{(c-v)\,\Delta t'} = \left(\frac{c}{c-v'}\right)\left(\frac{1}{\gamma_v \Delta t}\right) = \left(\frac{c}{c-v}\right)\left(\sqrt{1-\frac{v^2}{c^2}}\,\frac{1}{\Delta t}\right) = \sqrt{\frac{c+v}{c-v}}\; f(\text{moving toward each other}) \quad (2.3.3)$$

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:

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

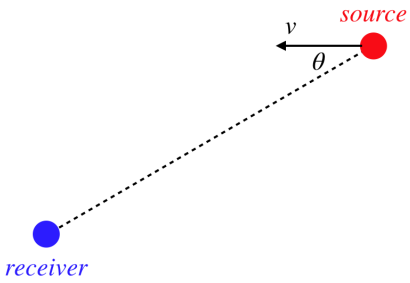
This effect of reducing the frequency perceived due to relative motion is called a *red shift*.

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 1.3.5 – 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 $v\cos\theta$:

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

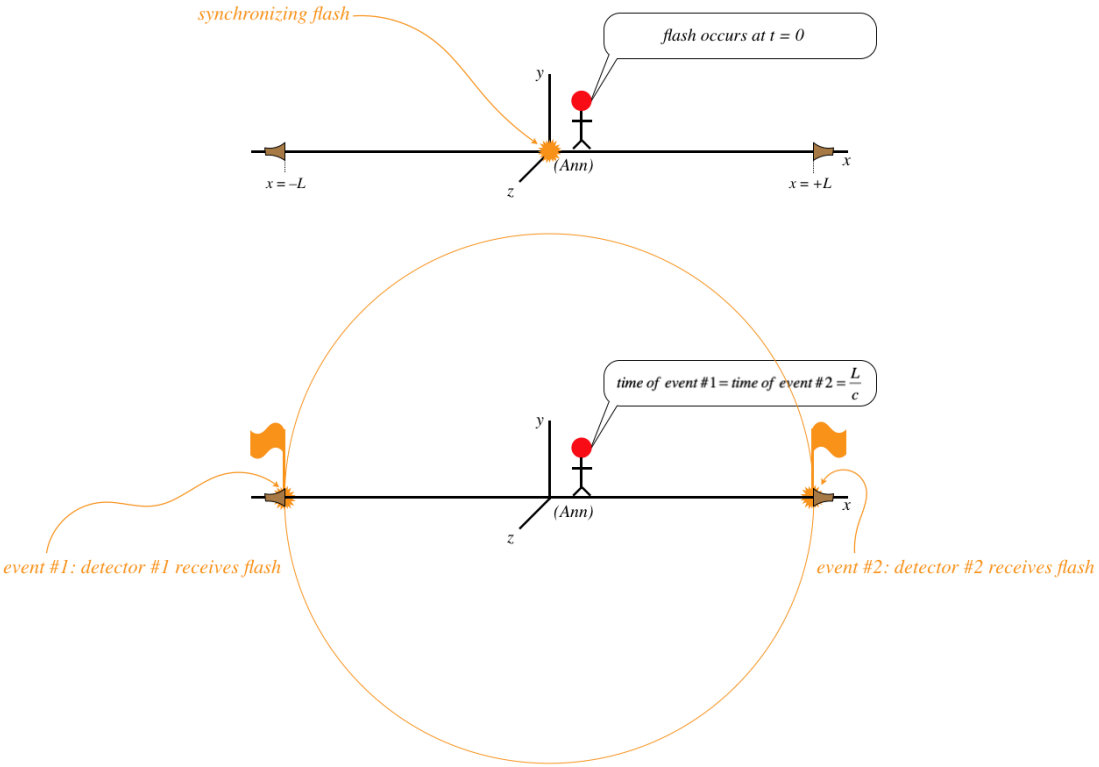
It is left as an exercise for the reader to show that this equation reduces to the three equations above for the appropriate values of θ.

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 fo far? Such a conundrum calls for a thought experiment!

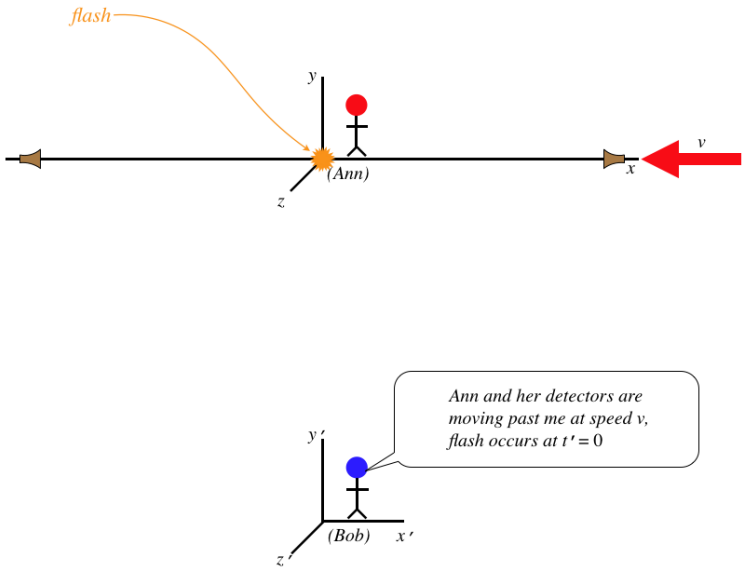
Let's suppose Ann decides to synchronize two clocks using a flash from her clock at the origin:

Figure 1.3.6 – Simultaneous Events for Ann



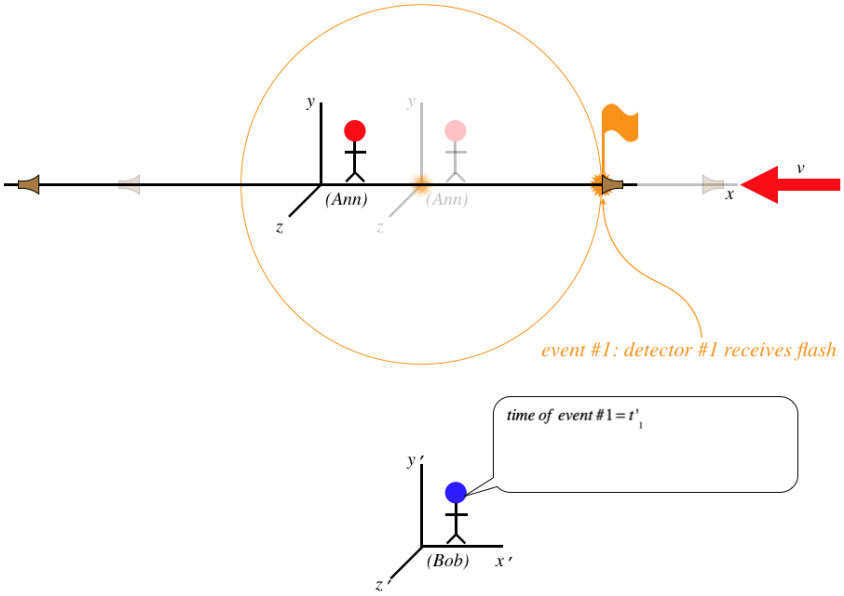
Does Bob agree that the two clocks (located at the two detectors) are synchronized? Let's look at what Bob sees:

Figure 1.3.7 – Ann's Synchronized Events Seen by Bob (a)



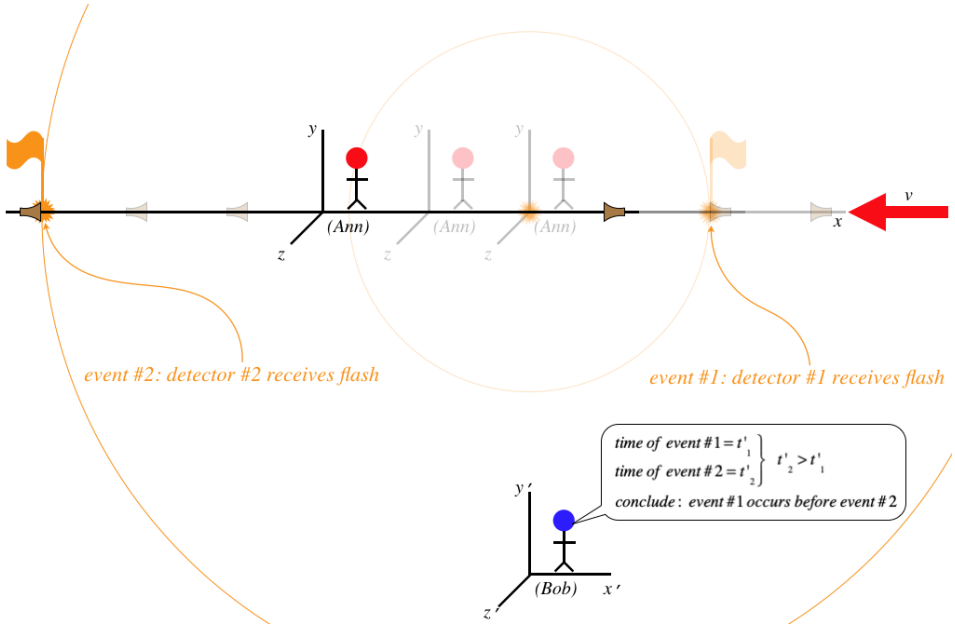
The detectors Ann is using are fixed in the lattice points in her frame, so they move along with her, according to Bob. When the light flashes, it takes time for the wave to get to the detectors, and while this time passes, the detectors move, according to Bob:

Figure 1.3.8 – 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 1.3.9 – 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 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:

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

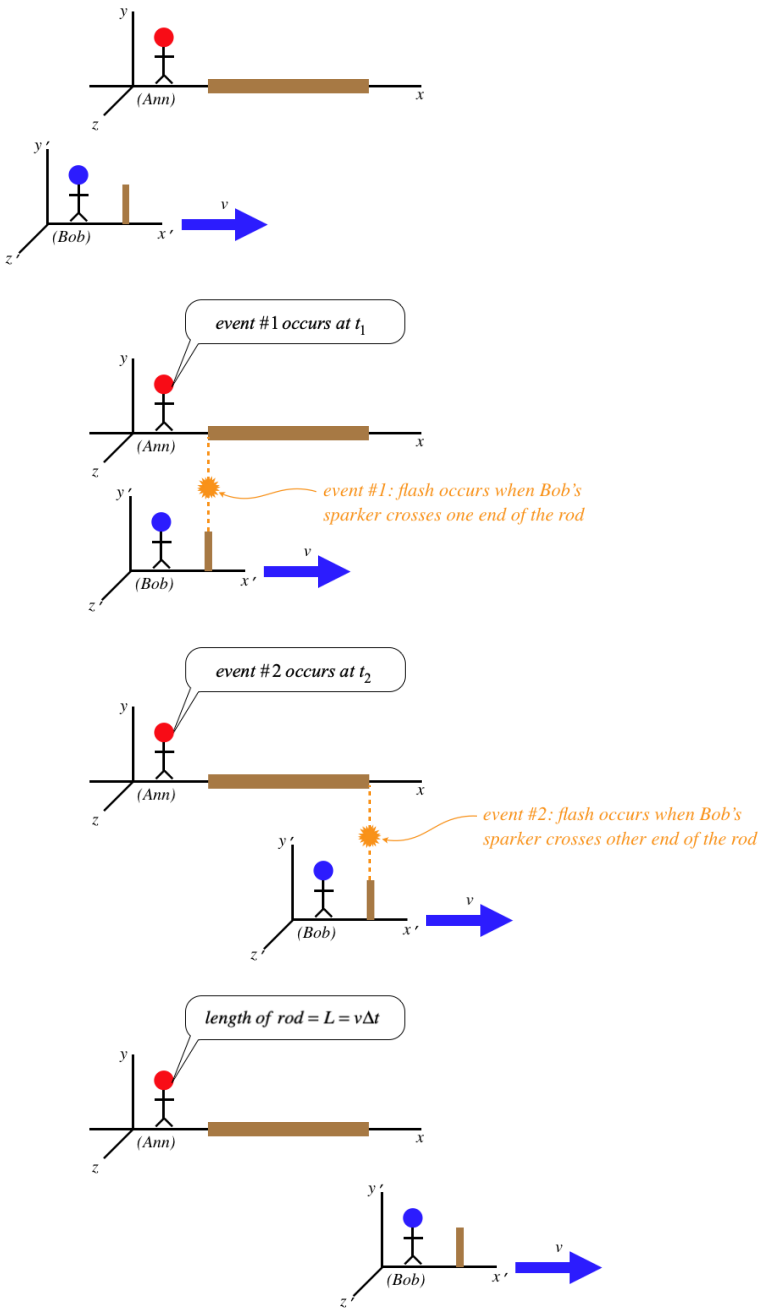
In the case above, the comparison is not between a coordinate time and a proper time, but two different coordinate times.

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.

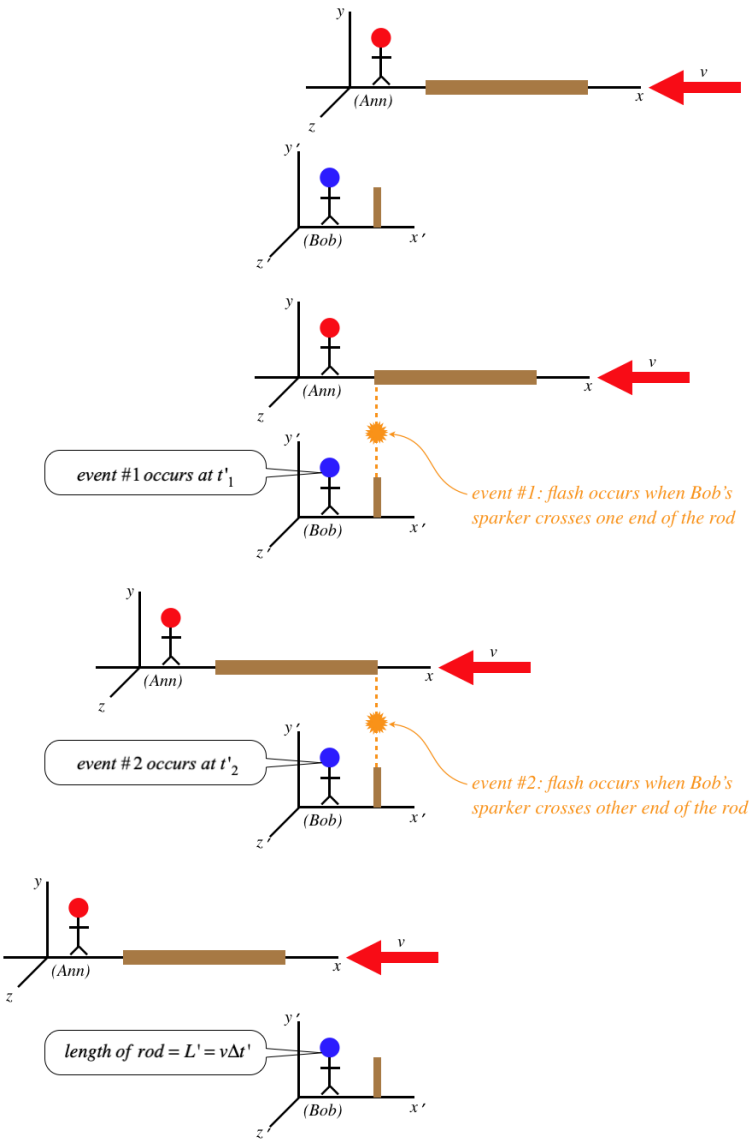
We consider The following scenario: Ann lays a rod down along her x-axis, and as she passes by Bob, each end of the rod creates a spark (constituting a spacetime event) when it coincides with a device stationary in Bob's frame that we will call a "sparker."

Figure 1.3.10 – 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 1.3.11 – 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:

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

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

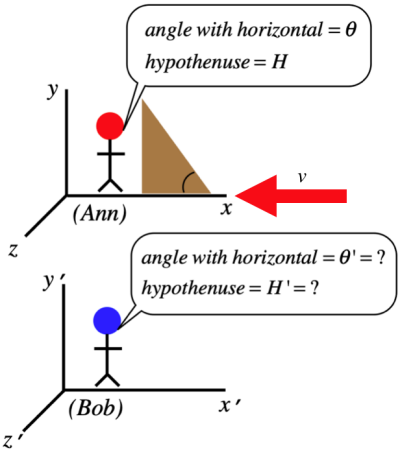
property inherent to objects, but length is significantly tougher to swallow. If it helps, it's okay to say that identical objects have equal proper lengths.

- 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 ate closer together along the x-direction that 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 1.3.12 – 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:

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

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:

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

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 is 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:

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

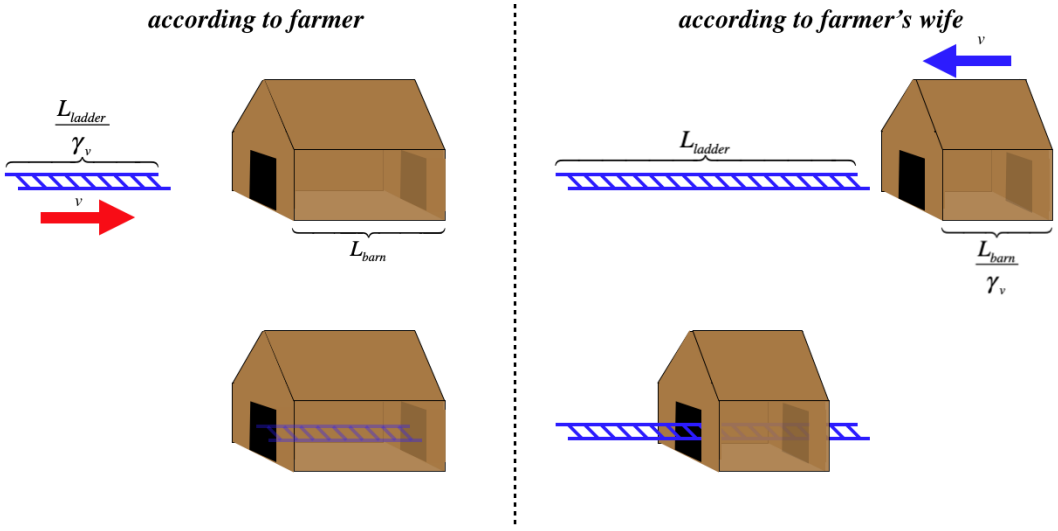
Since the $40\,ft$ barn is at rest relative to the farmer, the $40\,ft$ length-contracted ladder just barely fits.

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:

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

The frame of the farmer's wife sees the ladder at rest and the barn moving past, so naturally the $50\,ft$ ladder doesn't fit inside the $32\,ft$ length-contracted barn. So how it is 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?

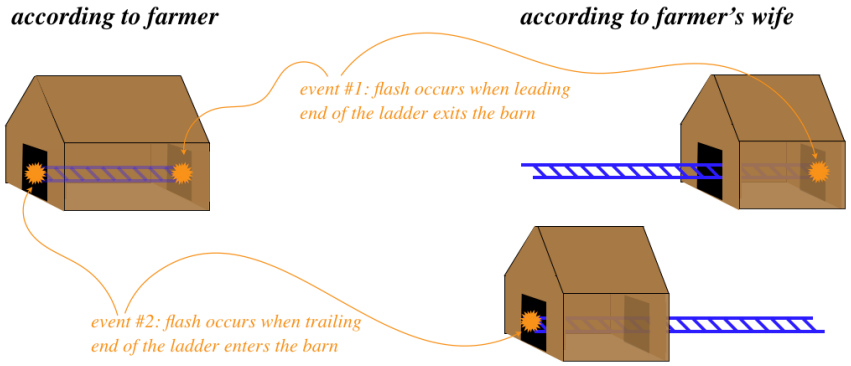
Figure 1.4.1 – Ladder & Barn Paradox



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.

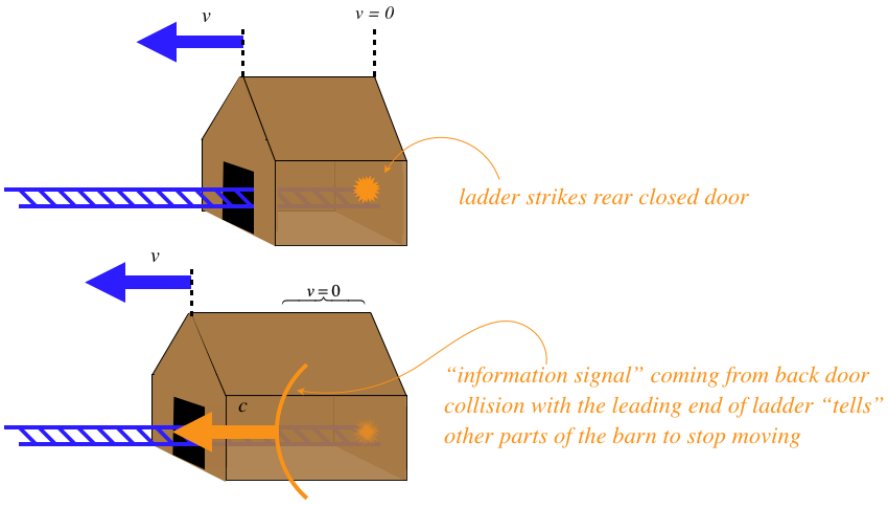
Okay, so much for the “dry run” of driving the ladder into one door and out the other. If the farmer can correctly claim that the ladder is within the barn, what happens if he closes the two barn doors simultaneously, while the ladder is in there? What will his wife see then?

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?



After t seconds in the wife's reference frame, the barn has gotten longer by an amount $v\Delta t$:

$$\text{length of barn after } t = 32ft + vt = 32ft + (0.6c)\,t \tag{2.4.3}$$

The distance that the light has traveled from the rear door in the wife's reference frame is equal to ct, so the signal reaches the front of the barn when:

$$ct = 32ft + (0.6c)\,t \quad \Rightarrow \quad t = \frac{32ft}{0.4c} \tag{2.4.4}$$

Okay, so we know how long it takes according to the wife for the barn to stop moving, so we can plug this time in to determine the maximum length of the barn:

$$\text{length of stretched barn} = \text{distance the light travels in the total time } = ct = \frac{32ft}{0.4} = 80ft \tag{2.4.5}$$

The ladder is 50ft long in the wife's frame, so it is inside the barn before the front of the barn gets the message to stop moving, and the barn door can be closed, averting the paradox entirely.

The Twin Paradox

The next paradox takes on time dilation. The idea is that if Ann and Bob can both see the other's clock ticking more slowly, what happens if, after awhile, we just bring both of the clocks together (into the same frame) and compare them? Which clock will have elapsed more time?

Two identical twins are born on a space station, and live there for their whole lives, until one of them gets into a spaceship for a long journey. When she speeds away from the space station at a sizable fraction of the speed of light, the twin on the space station (Ann) uses a powerful telescope to look through the window of her sister Bea's spaceship, and sees the effects of time dilation as the clock on Bea's ship is turning much slower than an identical clock in the space station. Bea also has a powerful telescope, and looking back at her sister, see that the space station clock is turning slow compared to her own. When Bea returns from her long journey and reunites with her sister, it would seem that both twins will encounter someone that is much younger than they are, but of course this is impossible. Which one will be older, or will they be the same age at the reunion?

The apparent paradox arises here because of symmetry. Each twin can claim that the other is moving because there is no such thing as a universal inertial frame. What is more, if we define the spacetime events as being the flash of light at the space dock that 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:

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

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:

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

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 our discussion around Equation 1.2.9 – the proper time measurement depends upon 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. So the integrals come out different for the two sisters. This appeal to pure mathematics is perhaps not very satisfying, so let's try a new thought experiment to try 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:

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

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):

$$\Delta\tau_{Chu} = \Delta\tau_{Chu}\,(leg\ 1) + \Delta\tau_{Chu}\,(leg\ 2) = \frac{\Delta t_{Ann}\,(leg\ 1)}{\gamma_{v_1}} + \frac{\Delta t_{Ann}\,(leg\ 2)}{\gamma_{v_2}} = \sqrt{1-0.25^2}\,(40min) \tag{2.4.9}$$

$$+\sqrt{1-0.75^2}\,(40min) = \frac{\sqrt{15}+\sqrt{7}}{4}(40min) \approx 65min$$

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).

CHAPTER OVERVIEW

3: Kinematics in Special Relativity

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

3.1: Spacetime Diagrams

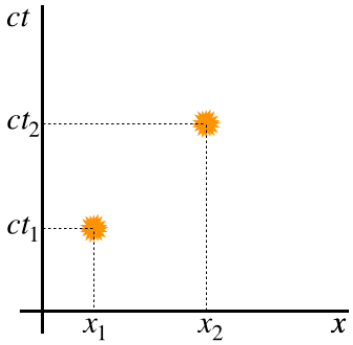
Graphing Spacetime Events

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



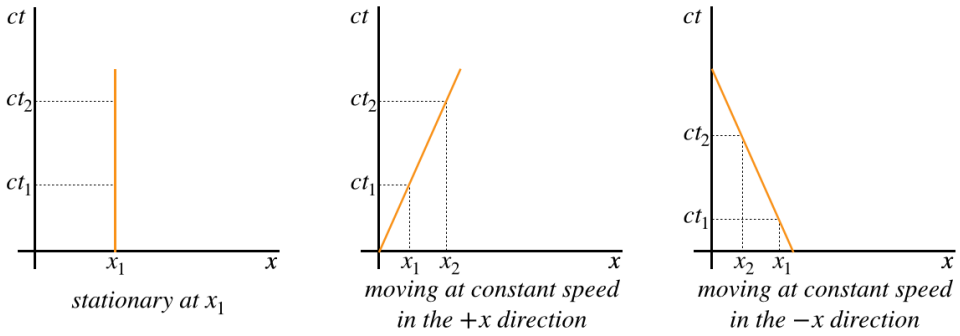
The times indicated by the values on the vertical axis are those measured by synchronized clocks in the frame of the observer represented by these axes. If the two events happen to be aligned vertically, then the two events occur at the same position in this frame, which means the difference $t_2 - t_1$ 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:

$$\text{slope of a world line} = \frac{ct_2 - ct_1}{x_2 - x_1} = \frac{c}{u}\ , \quad u \equiv \frac{\Delta x}{\Delta t} = \text{speed of the object measured by observer} \tag{3.1.1}$$

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.

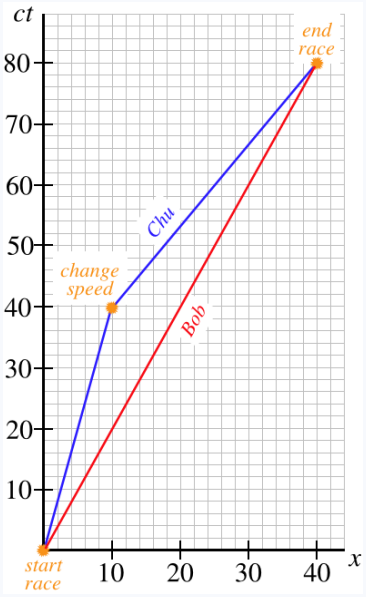
Example 3.1.1

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.

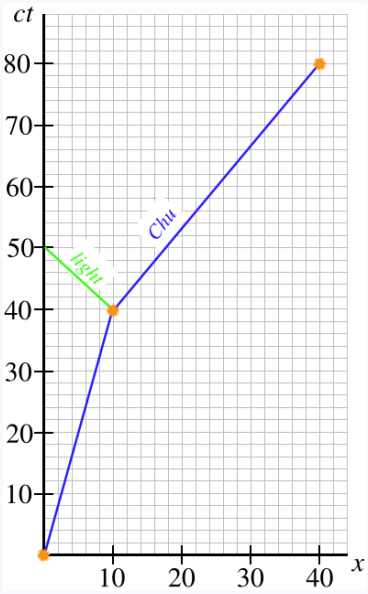
a. *Draw a spacetime diagram in Ann's reference frame depicting the world lines of both Bob and Chu, and label the important spacetime events along these worldlines.*
b. *Use the diagram to determine the time on Ann's clock in her spaceship (not at the lattice point in her reference frame) when she **sees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

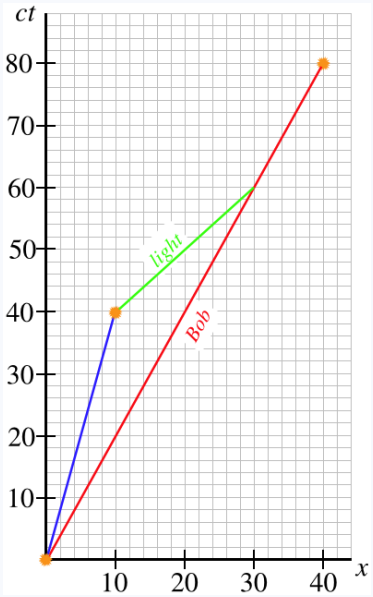


b. Ann doesn't see the event of Chu changing speed until light from that event reaches her. The world line of this light has a slope of -1 because it is coming back to Ann (in the $-x$-direction). Drawing this into the diagram shows that she discovers Chu has changed speeds at time $t = 50$ minutes, 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:

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



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.

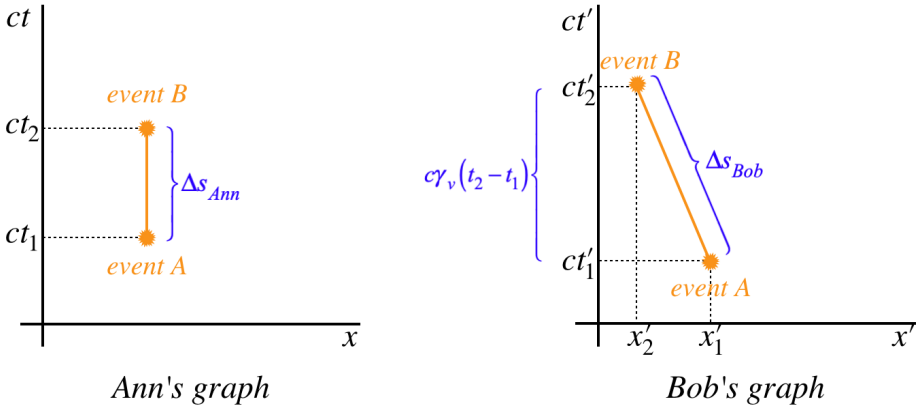
As we showed above, the spacetime interval between the two events is the *square of the length of the segment of the world line* connecting the two events in Ann's spacetime diagram. When we first discussed the nature of time, we made a point of noting that all proper times, and therefore the spacetime 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):

$$t'_2 - t'_1 = \gamma_v \left(t_2 - t_1\right) \tag{3.1.2}$$

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:

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

Eliminating the $2L$ from these equations gives:

$$c^2\Delta t^2 = c^2\Delta t'^2 - \Delta x'^2 \tag{3.1.4}$$

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:

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

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.

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

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:

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

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

$$\Delta s = \int_{\text{along world line}} ds = \int_{\text{along world line}} cdt\sqrt{1 - \frac{u^2}{c^2}} \tag{3.1.8}$$

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:

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

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 were 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:

a. *Compute the (Minkowski) length of the world line for Bob to get the time elapsed for him during the race.*
b. *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:

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

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.

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

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):

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

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.

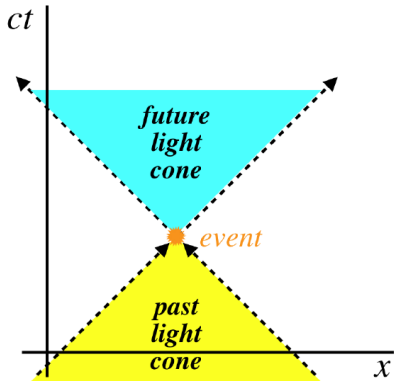
space-like separation: $\Delta s^2 < 0$

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.

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:

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

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:

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

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:

$$0 = J \cdot vt - \gamma_v vt \quad \Rightarrow \quad J = \gamma_v \tag{3.2.7}$$

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):

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

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

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

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 $\frac{v}{c} \to 0$ 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:

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

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:

$$c\Delta t' = \gamma_v \left[c\Delta t - \left(\frac{v}{c}\right) \cancelto{0}{\Delta x} \right] \quad \Rightarrow \quad \Delta t' = \gamma_v \Delta t \tag{3.2.10}$$

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

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

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:

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

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

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

3.3: Velocity Addition

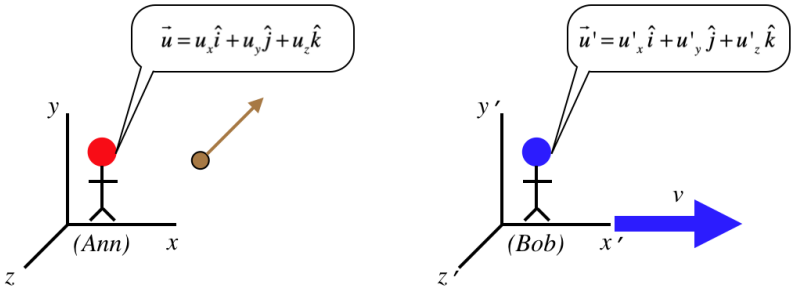
Cosmic Speed Limit

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:

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

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.*]

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

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 velcoity with Ann gives the speed measured by Bob:

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

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:

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

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:

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

3.4: Electricity and Magnetism

Motion in E&M

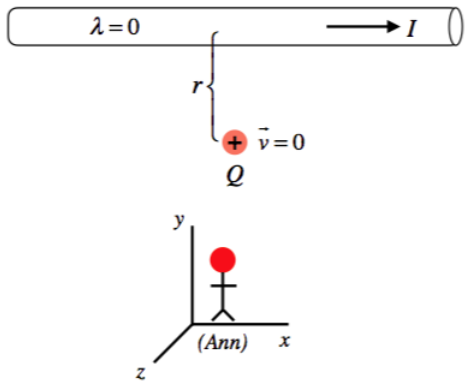
Electricity and magnetism was the main wildly-successful theory at the time that Einstein came onto the scene. Maxwell had shown how the two different forces could be unified into one mathematical theory. Electric charges cause both electric and magnetic fields, both of the fields exert forces on electric 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:

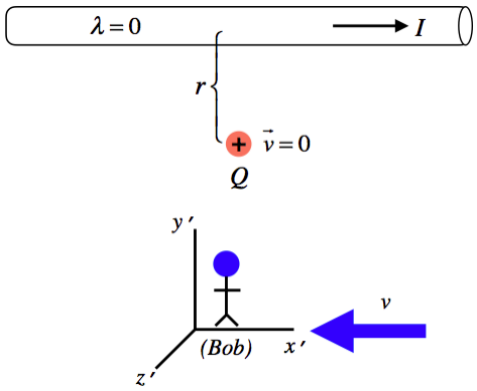
Figure 3.4.1 – Ann Observes a Stationary Charge Near a Current-Carrying Wire



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:

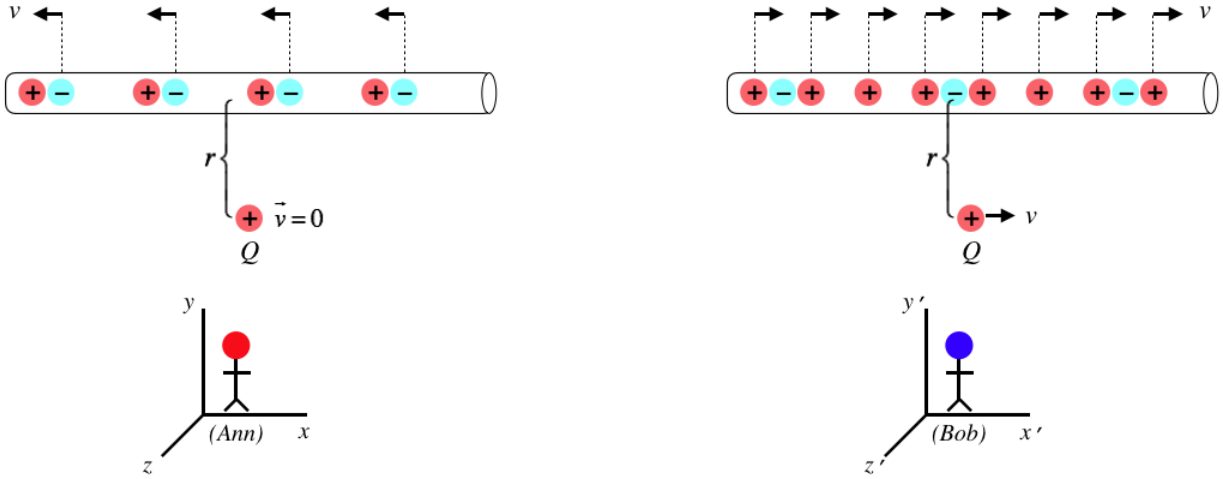
Figure 3.4.2 – Bob Moves by the Charge with the Electrons



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 see 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).

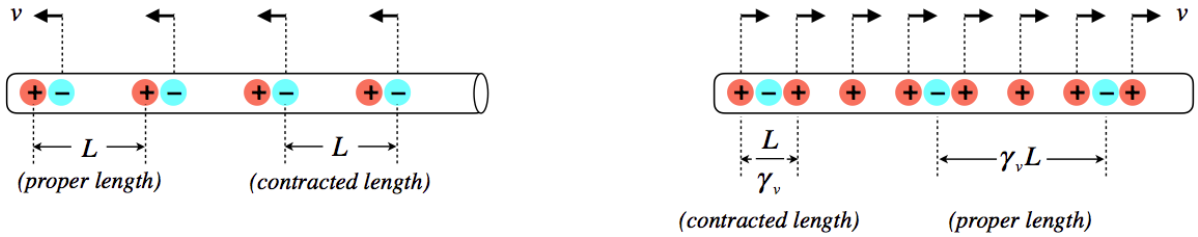
Without yet doing any math, we can begin to see why Bob's conclusion could possibly agree with Ann's. With a current to the right and a right-moving point charge, as we noted above, Bob does witness a magnetic force on the point charge toward the wire. But now we see that there is also an electric force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

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

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:

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

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

$$I = \frac{dq}{dt} = \frac{dq}{dx}\frac{dx}{dt} = \lambda v \tag{3.4.3}$$

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:

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

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:

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

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:

$$\vec{F} = Q\vec{E} + Q\vec{v}\times\vec{B} \quad \Rightarrow \quad F_{Bob} = Q\left(\frac{v^2}{c^2\epsilon_o}\right)\left(\frac{\gamma_v q}{2\pi r L}\right) - Qv(\mu_o v)\left(\frac{\gamma_v q}{2\pi r L}\right) = \left(\frac{1}{c^2\epsilon_o} - \mu_o\right)\left(\frac{\gamma_v v^2 qQ}{2\pi r L}\right) \tag{3.4.6}$$

And now the calculation is completed at last using another result from Physics 9C:

$$c^2 = \frac{1}{\mu_o\epsilon_o} \quad \Rightarrow \quad F_{Bob} = 0 \tag{3.4.7}$$

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:

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

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 $\vec{v}$.

CHAPTER OVERVIEW

4: Dynamics in Special Relativity

4.1: Momentum Conservation

4.2: Energy Conservation

4.1: Momentum Conservation

Another Thought Experiment

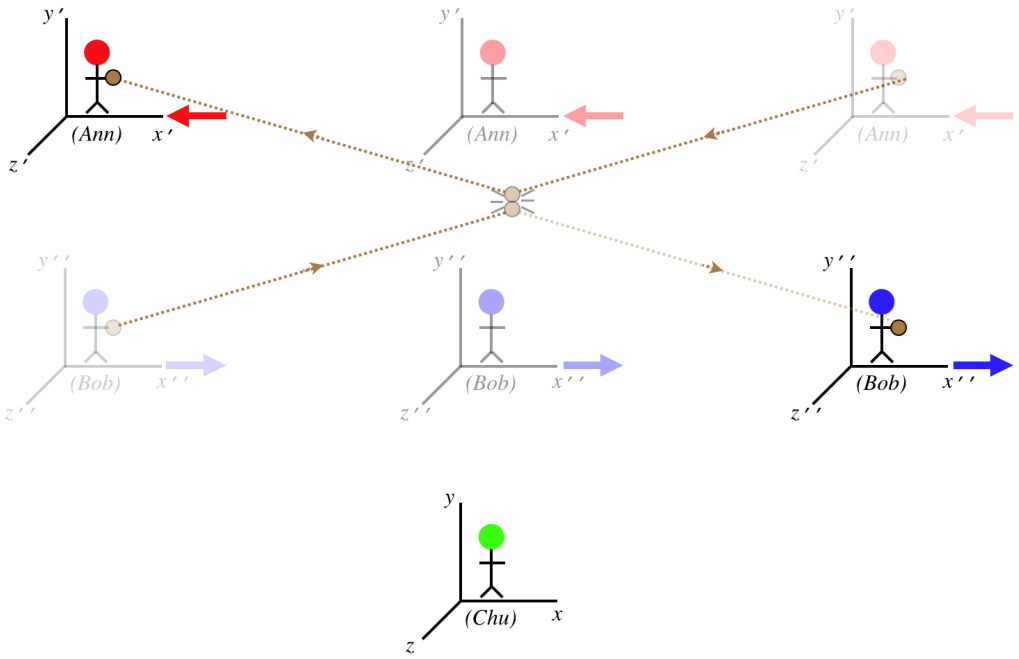
We have completed our exploration of the fundamentals of relativistic kinematics and its consequences. Now it is time to examine the consequences of the relativity principle in the area of dynamics. Our first clue that something needs to be done differently comes from the basis of all dynamics: Newton's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

$$\vec{F}_{net} = \frac{d\vec{p}_{cm}}{dt} \tag{4.1.1}$$

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!

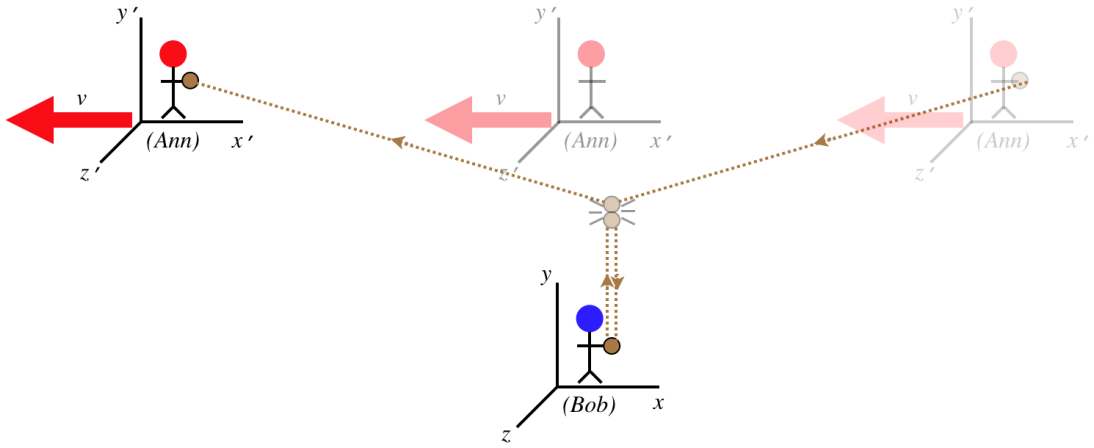
We have three players involved here (Ann, Bob, and Chu), and we are initially viewing the activities from Chu's perspective, who sees Ann moving in the $-x$-direction and Bob moving in the $+x$-direction at equal speeds. Ann and Bob both throw identical balls in a direction that is (from their own perspectives) 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):

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

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 is 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 principal 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:

$$\vec{p} \equiv m\gamma_u \vec{u} \tag{4.1.3}$$

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 $\gamma_{u'}$ 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 $u << c$, $\gamma_u \to 1$.

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:

$$\overrightarrow{p}_A = -m\,\gamma_{uA}\,u_A\,\hat{j} = -\frac{m\,u}{\sqrt{1-\frac{u^2}{c^2}}}\,\hat{j} \tag{4.1.4}$$

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:

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

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:

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

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):

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

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:

$$\frac{1}{\sqrt{1-\frac{v^2}{c^2}-\frac{u^2}{\gamma_v^2 c^2}}}\,\cancel{m}\,\frac{\cancel{u}}{\gamma_v} = \frac{m\,\cancel{u}}{\sqrt{1-\frac{u^2}{c^2}}} \tag{4.1.8}$$

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

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

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.

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.

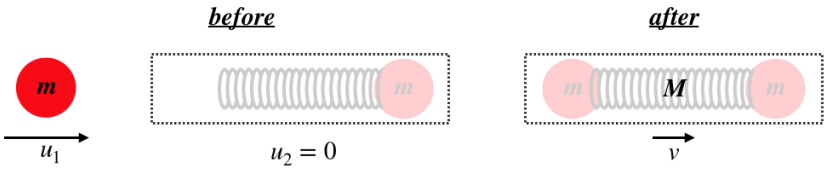
In Physics 9A, for the non-relativistic case of a collision of this kind, we showed that the fraction of the initial kinetic energy the box system has after the collision is given by:

$$\frac{KE_{after}}{KE_{before}} = \frac{m_1}{m_1 + m_2} = \frac{1}{2} \tag{4.2.1}$$

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.

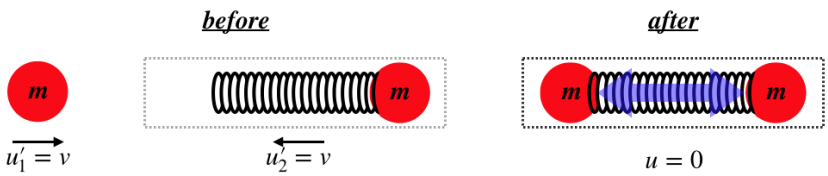
Let's see how all this works out for relativity with our new definition of momentum. We will watch this collision from two different perspectives. The first is Ann, who sees the collision from the perspective shown above, and who cannot see inside the boxed system, so she doesn't even know the system's 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:

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

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

$$u_1 = \frac{u_1' + v}{1 + \frac{u_1' v}{c^2}} = \frac{2v}{1 + \frac{v^2}{c^2}} \quad \Rightarrow \quad \frac{u_1}{v} = \frac{2c^2}{c^2 + v^2} \tag{4.2.3}$$

Writing γ_{u_1} in terms of v:

$$\gamma_{u_1} = \frac{1}{\sqrt{1 - \frac{u_1^2}{c^2}}} = \frac{1}{\sqrt{1 - \left(\frac{u_1}{v} \cdot \frac{v}{c}\right)^2}} = \frac{1}{\sqrt{1 - \left(\frac{2c^2}{c^2+v^2} \cdot \frac{v}{c}\right)^2}} = \frac{c^2 + v^2}{\sqrt{c^4 - 2v^2c^2 + v^4}} = \frac{c^2 + v^2}{c^2 - v^2} \tag{4.2.4}$$

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

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

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:

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

Now using the usual $v << c$ approximation:

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

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:

$$\text{internal energy} = \frac{1}{2}Mv^2 = \Delta m c^2 \tag{4.2.8}$$

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:

$$E_{rest} = Mc^2 \tag{4.2.9}$$

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:

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

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:

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

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:

$$E_{tot} = \gamma_u mc^2 \tag{4.2.12}$$

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:

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

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:

$$\frac{c^2 + v^2}{c^2 - v^2} + 1 = \frac{2}{1 - \frac{v^2}{c^2}}$$

$$\frac{c^2 + v^2 + c^2 - v^2}{c^2 - v^2} = \frac{2c^2}{c^2 - v^2}$$

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:

$$KE = (\gamma_u - 1)mc^2 \tag{4.2.13}$$

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:

$$\begin{aligned} KE &= (\gamma_u - 1)mc^2 \\ \frac{p^2}{2m} &= \tfrac{1}{2m}(\gamma_u mu)^2 = \tfrac{1}{2}\gamma_u^2 mu^2 \end{aligned} \tag{4.2.14}$$

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

$$E^2 = \gamma_u^2 m^2 c^4 = \frac{c^2}{c^2 - u^2} m^2 c^4 = \left(\frac{u^2}{c^2 - u^2} + 1\right) m^2 c^4 = \frac{c^2}{c^2 - u^2} m^2 u^2 c^2 + m^2 c^4 = \gamma_u^2 m^2 u^2 c^2 + m^2 c^4 = p^2 c^2 + m^2 c^4 \tag{4.2.15}$$

So the alternative ways of writing the total energy are:

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

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:

$$E_{photon} = pc \tag{4.2.17}$$

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.

CHAPTER OVERVIEW

5: Light as a Particle

5.1: Blackbody Radiation
5.2: The Photoelectric Effect
5.3: Compton Effect
5.4: Double-Slit Experiment

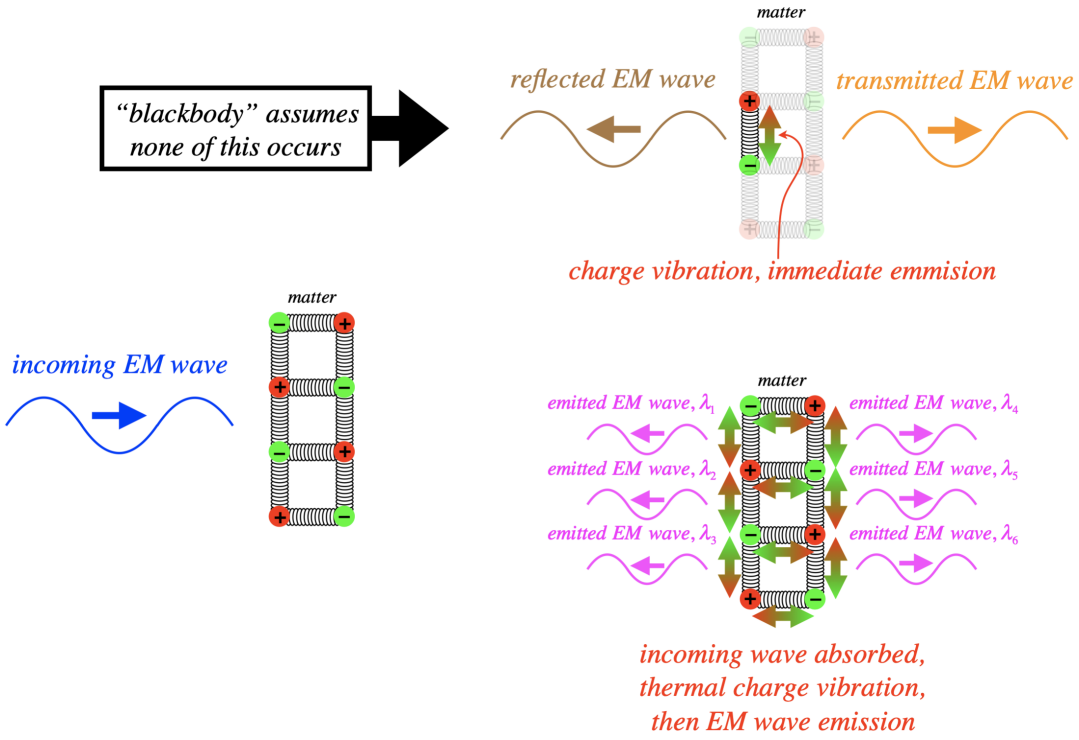
5.1: Blackbody Radiation

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

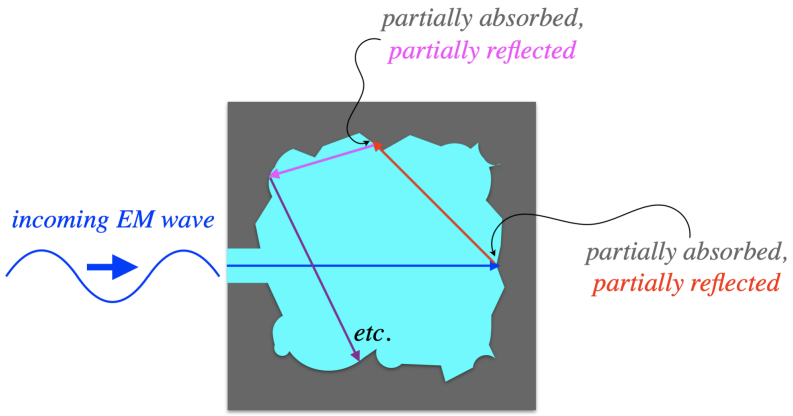


One might ask where the name "blackbody" comes from. Imagine some yellow light striking such an object. If it is totally reflected, then yellow light is emitted by it. But if the light's energy is absorbed by the object, and that energy is distributed randomly through the atoms which then emit light 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 specturm 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



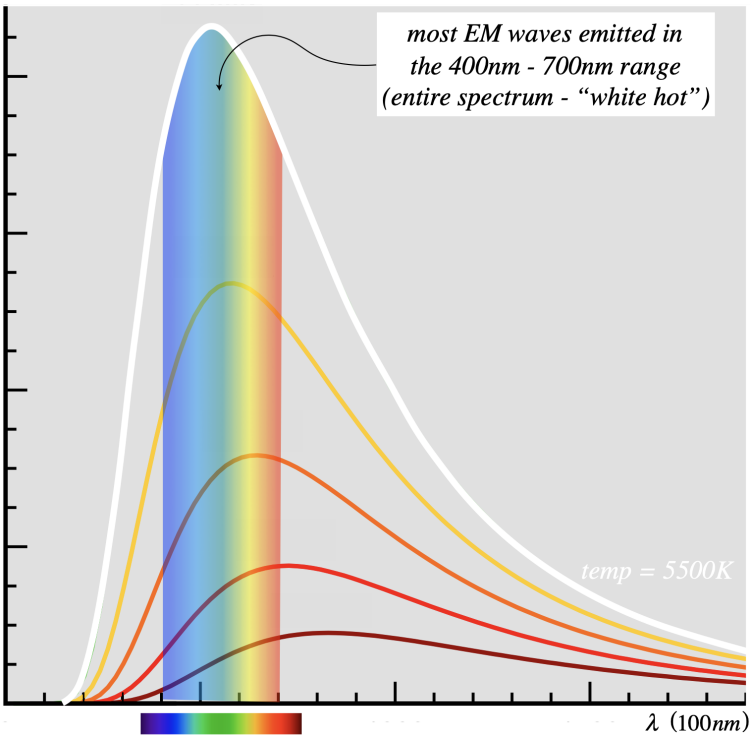
In the late 1800's, experiments were performed with these makeshift blackbodies. Note that since the charged particles are moving thermally, their motions can have any of a wide range of frequencies (and therefore so can the EM waves they emit). Physicists were interested in explaining how the outgoing 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:

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

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:

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

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:

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

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 is h 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:

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

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.

5.2: The Photoelectric Effect

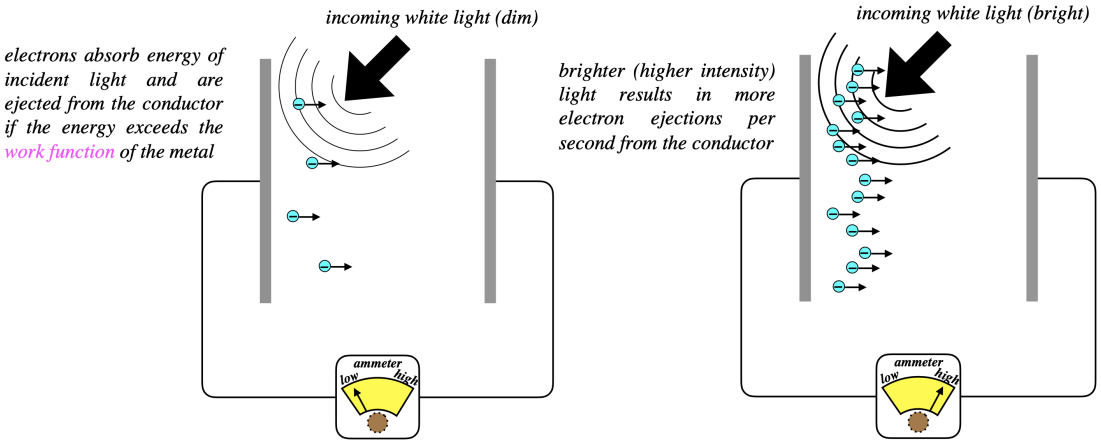
Light Interacting with Conductors

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 form the battery, the electron flow can be measured by an ammeter. If we turn up the brightness of the light, the measured current rises.

Figure 5.2.1 - Photoelectric Effect (Unsurprising)



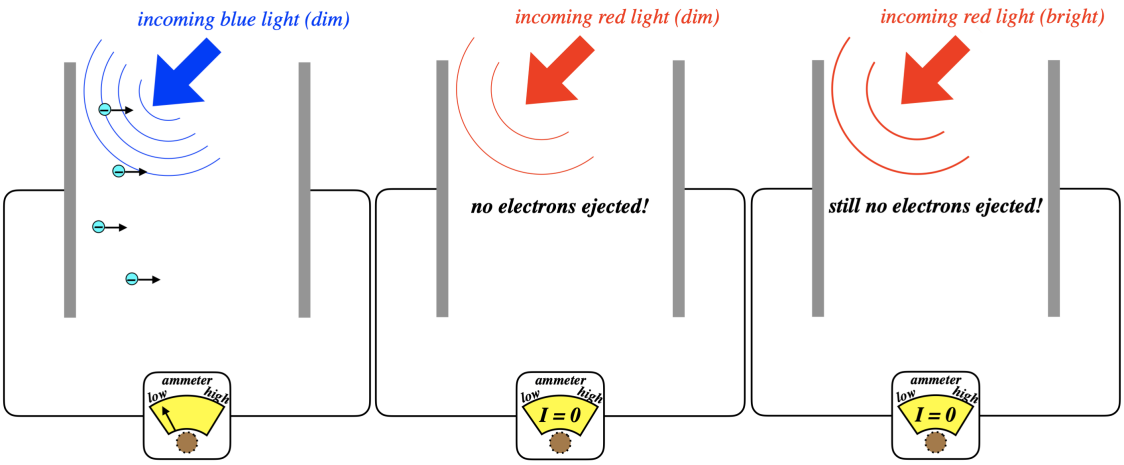
Digging Deeper

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



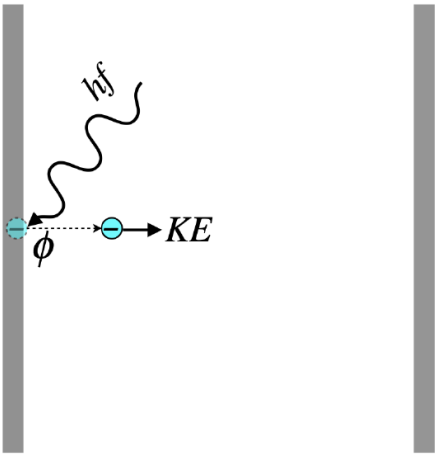
The blue light ejects electrons even when dim, as the white light did, but dim red light does not. This tells us that it was the blue part of the spectrum that was ejecting the electrons when the white light was shone on the metal earlier. While this result is peculiar from our standard understanding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

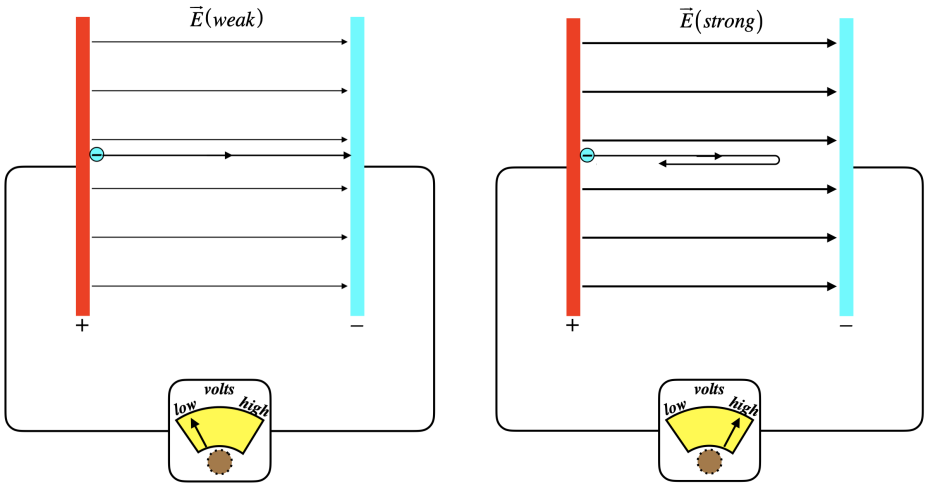
This answers the effect of frequency, but what about the second "digging deeper" question regarding the energy of the electrons that are ejected? Einstein's solution gives us that answer as well. Applying conservation of energy to this process gives us immediately what we seek:

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:

$$e\Delta V_{stopping} = KE_{max} = hf - \phi \tag{5.2.1}$$

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.

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:

$$E_{photon} = pc \tag{5.3.1}$$

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 $theta$. Calling the momentum of the incoming photon p_{in}, the momentum of the outgoing momentum p_{out}, and using the relativistic momentum for the particle, we have the following two conserved momentum components:

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

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

$$p_{in}c + mc^2 = p_{out}c + \gamma_u mc^2 \tag{5.3.3}$$

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

$$\frac{1}{p_{out}} - \frac{1}{p_{in}} = \frac{1}{mc}(1 - \cos\theta) \tag{5.3.4}$$

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

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

Putting this in above gives:

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

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, $\cos\theta = \cos 180^o = -1$, which means that the wavelength of the incoming light is increased by two Compton wavelengths.

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

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…

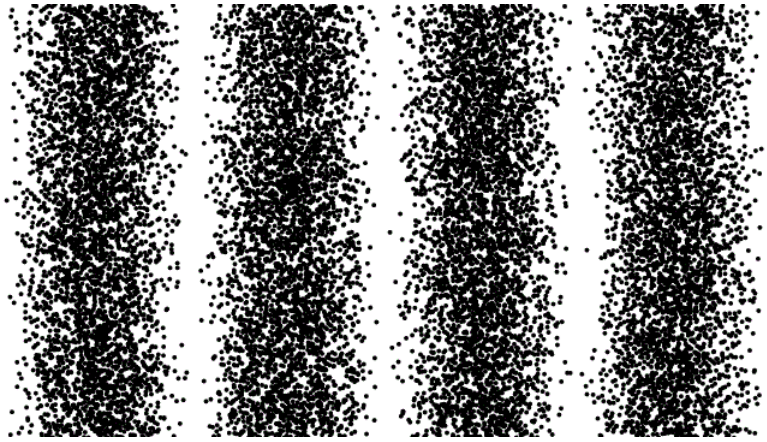
Light (a wave) arrives in phase at two slits, and part of the wave originated from each slit. The light (a wave) from each slit source spreads out to all parts of the screen and at each point on the screen the two waves arrive at some relative phase. If the phases are different by an even integer times 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 n 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".

CHAPTER OVERVIEW

6: Matter as a Wave

6.1: From Light to Electrons
6.2: Interpreting Matter Waves

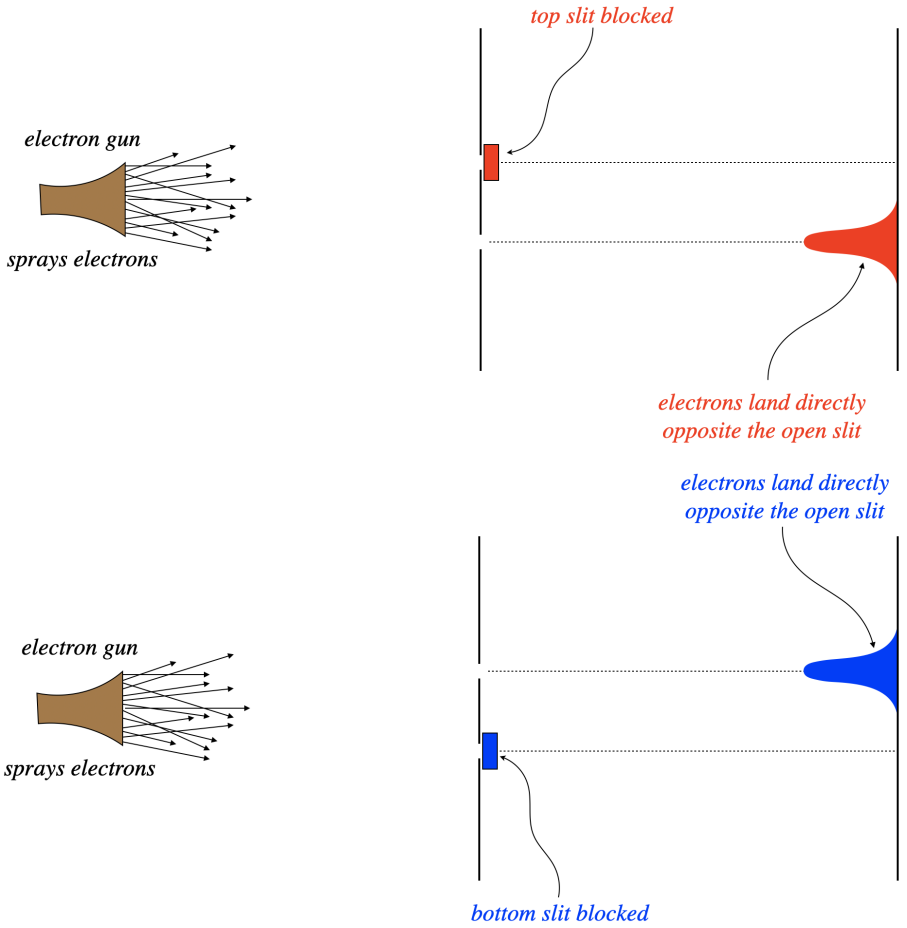
6.1: From Light to Electrons

The Davisson-Germer Experiment

Light's apparent dual nature as a wave and a particle might lead one to wonder, "If something we always thought was just a wave has particle properties, might not things we always thought to be particles have wave properties?" In our discussion of Compton scattering, we made an association between momentum and wavelength for light. Well, clearly particles like electrons have momentum, so maybe they have equivalent wavelengths, and therefore wave behavior. We'll explore this possibility with the results of series of experiments, accompanied by a particularly enlightening narrative first delivered by the late Richard Feynman.

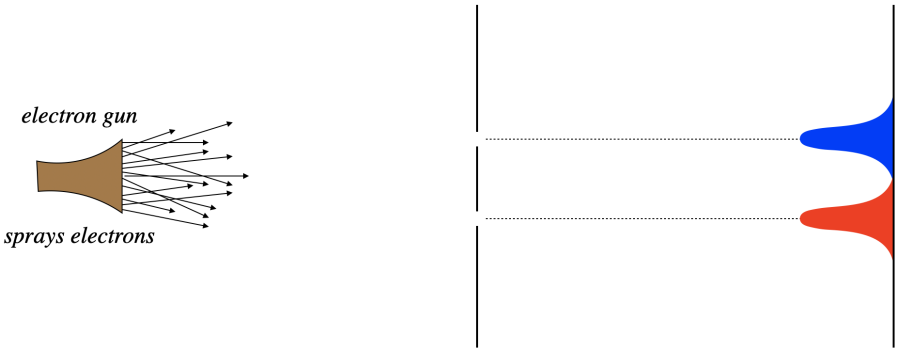
Consider firing electrons at a double-slit apparatus, with one of the two slits blocked. With one path to the screen, we see a distribution of electron-caused dots on the screen exactly as we would expect – in a cluster centered directly opposite the slit. Next the slits are reversed – the previously-closed 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 6.1.1 – Electrons Through a Double Slit with One Slit Blocked



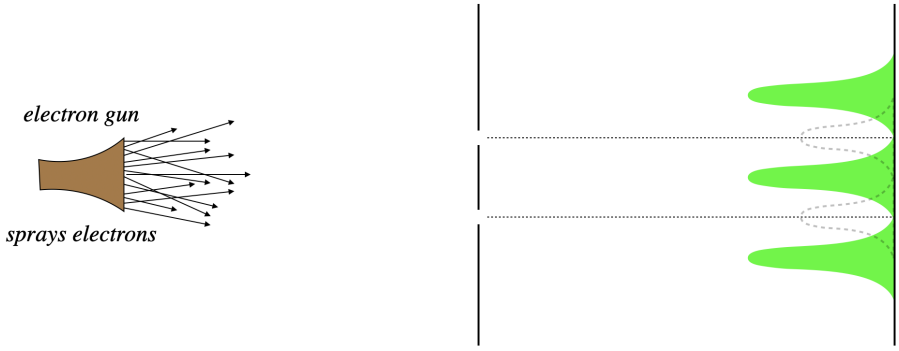
Given these results, the natural prediction of what happens when both slits are open is that we see both of the patterns we saw previously, at the same time. All the electrons that pass through the top slit should end up across from it, and all of those that pass through the bottom slit end up across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

Figure 6.1.2 – Electrons Through a Double Slit with Both Slits Open (Expected)



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

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

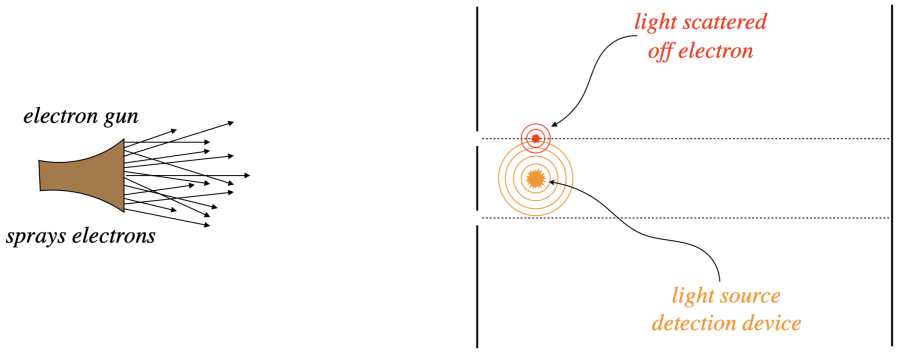


The first attempt to explain this is naturally to say that the electrons are interacting with each other after they pass through the slits. But like the case of photons discussed in the previous section, the interference pattern emerges even if the electrons are sent through one at a time, once all the electron landings have occurred.

So the puzzling question is, "How does an electron, while going through the bottom slit, know whether or not the top slit is open?" If it is open, it's destination is different than if it is closed, but how can it tell where is 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. But that wave is not like any other wave, in that it carries information about the *probabilities* of landing at various points on the screen. This "probability wave" has the usual wave property that it can interfere with itself, creating some positions of zero probability, and others of relatively high probability, and these probabilities reflect the populations of the electron landing dots.

In an attempt to solve this puzzle, 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 6.1.4 – 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.

According to Compton scattering, the scattered light always has a longer wavelength than the light sent in, so we will be looking at very long wavelength light coming from the electron flashes. The flashes are separated by a distance roughly equal to the slit separation, and we find that the effect of the light on the interference pattern starts to diminish when the wavelength of the light we use gets longer than the slit separation. But then something new becomes a problem for our experiment. To determine where the light flash occurs, we need the light to carry with it some information about the distance scale – the wavelength of the light is a rough measure of the uncertainty of the position of its source. Just as we make the wavelength long enough for the interference pattern to return, it becomes too long to distinguish which slit the electron goes through. Infuriating... and amazing.

Comparing "Matter Waves" and Light Waves

We already have a mathematical description for light waves – Maxwell's equations. These don't explain why photons also act like particles, but one step at a time! Despite sharing the property of wave-particle duality, electrons and photons have some distinct differences, the two most notable being mass and electric charge. To keep these straight, we will employ the rather poor name of *matter waves* to describe the waves for electrons (and other particles), to distinguish them from light waves.

The fact that we get the same result for electrons as for light tells us that matter waves for free electrons (i.e. not under the influences of forces, like from electric or magnetic fields) must look very similar to those for light waves. For light, we found (by utilizing our Physics 9B wave function skills in Physics 9C), the following expression for the time and position dependence of the electric field magnitude for a light wave moving in the $+x$-direction (we'll choose phase equal to zero at $x = 0$, $t = 0$:

$$E(x,t) = E_o \cos\left[\frac{2\pi}{\lambda}x - \frac{2\pi}{T}\right] \tag{6.1.1}$$

Since we now know how this wave relates to physical properties of the photon (i.e. its momentum and energy), let's rewrite it in terms of those quantities:

$$p = \frac{h}{\lambda},\quad E = hf = \frac{h}{T} \quad\Rightarrow\quad E(x,t) = E_o \cos\left[\left(\frac{2\pi}{h}p\right)x - \left(\frac{2\pi}{h}E\right)t\right] \tag{6.1.2}$$

The quantity \(\frac{h}{2\pi} comes up frequently in quantum physics, so it is given its own symbol: $\hbar$ (pronounced "h-bar"). This makes our expression for the electric field:

$$E(x,t) = E_o \cos\left[\frac{p}{\hbar}x - \frac{E}{\hbar}t\right] \tag{6.1.3}$$

We know that the matter waves and light wave functions should have the same form, because when it comes to interference, waves are waves. So we expect that matter wave functions at least look *similar* to this. But what about the wave *equation*? We know that Maxwell showed that light waves can be described by the "usual" wave equation (again, simplifying to a wave traveling along the *x*-axis):

$$\frac{\partial^2 E}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} \tag{6.1.4}$$

It's easy to show that this wave equation works perfectly for light, because when we plug in our wave function from above, we get

$$\left.\begin{aligned}\frac{\partial^2 E}{\partial x^2} &= -\left(\frac{p}{\hbar}\right)^2 E(x,t)\\ \frac{\partial^2 E}{\partial t^2} &= -\left(\frac{E}{\hbar}\right)^2 E(x,t)\end{aligned}\right\}\quad \frac{\partial^2 E}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} \quad\Rightarrow\quad -\left(\frac{p}{\hbar}\right)^2 = -\frac{1}{c^2}\left(\frac{E}{\hbar}\right)^2 \quad\Rightarrow\quad E = pc \tag{6.1.5}$$

But as we saw in relativity, matter is very different from light – it has mass, and therefore this relationship between energy and momentum doesn't hold for matter. That means we need a different wave equation for matter waves than for light waves.

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 particles we will be dealing with will be moving at speeds that are not relativistic, and we'll use the Physics 9A-level relationship between kinetic energy and momentum:

$$KE = \frac{p^2}{2m} \tag{6.1.6}$$

Matter waves are not waves in electric and magnetic fields, so we need a symbol to describe the wave function, and the standard choice for this is the Greek letter *psi*: ψ. So 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 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-squared), 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. And finally, we need a factor of $2m$ introduced into the denominator to construct the kinetic energy/momentum relation. So let's try this:

$$\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) = \hbar\frac{\partial}{\partial t}\psi(x,t) \tag{6.1.7}$$
$$\text{where: } \psi(x,t) = \psi_o \cos\left[\frac{p}{\hbar}x - \frac{E}{\hbar}t\right]$$

The chain rules for the derivatives all work out nicely, but this wave equation falls short in the derivative itself – the derivative of cosine is (negative) sine, so while all the constants work out fine in front of the trig function, the trig functions themselves don't match.

A guy named Erwin Schrödinger didn't give up when he got this close. He realized that just as light waves have two parts (electric and magnetic), so too should matter waves. Here's how he incorporated two parts to the wave function: He allowed it to be a *complex number*. The real part of the wave function would be one part of the matter wave, and the imaginary part another. And just like for EM waves where changing electric fields give rise to magnetic fields and vice-versa, the real and imaginary parts of this wave function also mix. His solution is now known as *Schrödinger's equation* (for a free particle):

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) = i\hbar\frac{\partial}{\partial t}\psi(x,t) \tag{6.1.8}$$
$$\text{where: } \psi(x,t) = \psi_o \cos\left[\frac{p}{\hbar}x - \frac{E}{\hbar}t\right] + i\sin\left[\frac{p}{\hbar}x - \frac{E}{\hbar}t\right]$$

We can shorten the formula for the wave function by using a famous identity attributed to Euler:

$$e^{i\theta} = \cos\theta + i\sin\theta \quad \Rightarrow \quad \psi(x,t) = \psi_o e^{i\left(\frac{p}{\hbar}x - \frac{E}{\hbar}t\right)} \tag{6.1.9}$$

6.2: Interpreting Matter Waves

de Broglie Wavelength

We have described matter waves in terms of the momentum and energy of the particle, but it is still a wave, and as such possesses wave properties like wavelength and frequency. We can extract these quantities directly from the wave function by comparing it to a wave function of a general harmonic wave. We'll start with the matter wave's wavelength, known as the *de Broglie wavelength*:

$$\frac{2\pi}{\lambda} = \frac{p}{\hbar} \quad \Rightarrow \quad \lambda = \frac{h}{p} \tag{6.2.1}$$

This is the wavelength we need to use in (for example) a double slit calculation to predict interference patterns. Notice that unlike the photon, this wavelength depends upon the mass of the particle and its speed.

An Application

Obviously observations in science are highly-dependent upon light – telescopes collect light from outer space, and microscopes collect light from very small dimensions, for example. The microscope in particular runs into a limitation that comes from the wave nature of light. In the discussion in the previous section of trying to watch electrons as they go from the double slit to the screen, we said that the ability of the light to resolve the location of the electron depends on the wavelength of the light – the shorter the wavelength, the finer the granularity of the resolution.

So what if we use electrons for imaging, rather than light? They behave like waves like light does, but they have other features that light doesn't have, such as our ability to alter their speed and direction with electric and magnetic fields (so for example, we can focus them with magnetic fields to achieve the same effect as focusing light with a lens). But the real kicker is their resolving power thanks to their short de Broglie wavelengths. Such a device is known as an electron microscope. Let's do the math to see why these work so well...

Imagine accelerating a stream of electrons through a voltage of say 1000 volts (CRT televisions accelerated electrons with voltages over 10 times this great, and the best electron microscopes are much higher still). We can calculate the wavelength (and therefore the resolving power) of the matter waves thus created:

$$KE = q_e \Delta V = \left(1.6 \times 10^{-19} C\right)\left(1000 V\right) = 1.6 \times 10^{-16} J$$

$$\frac{p^2}{2m_e} = KE \quad \Rightarrow \quad p = \sqrt{2\left(9.11 \times 10^{-31} kg\right)\left(1.6 \times 10^{-19} C\right)} = 1.71 \times 10^{-23} \frac{kg \cdot m}{s}$$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34} J \cdot s}{1.71 \times 10^{-23} \frac{kg \cdot m}{s}} = 3.88 \times 10^{-11} m$$

This is on the atomic scale! If we wanted to use light to probe the same dimensions, we would need to use some serious X-rays (not the garden-variety dentist chair kind), and the energy per photon to achieve this wavelength would be:

$$E_{photon} = hf = \frac{hc}{\lambda} = \frac{\left(6.63 \times 10^{-34} J \cdot s\right)\left(3.0 \times 10^{8} \frac{m}{s}\right)}{3.88 \times 10^{-11} m} = 5.13 \times 10^{-15} J$$

This is more than 30 times greater than the $1.6 \times 10^{-16} J$ of energy required by an electron, for the same result!

Probability Density

After associating the wave property of wavelength with a physical property (momentum), and the wave property of frequency with the physical property of energy, we turn our attention to wave amplitude. With light waves, we know that the amplitude-squared is a measure of intensity. When we first discussed wave-particle duality, we decided that in a particle context, intensity at a given position is a measure of the number of particles arriving at that position per second. With our one-photon-per-hour double-slit experiment, we concluded that this particle arrival rate is really a measure of the *probability* of a single particle arriving at that position.

Of course, the probability of arriving at a specific exact position is zero, since the sum of all the probabilities has to equal 1, and there is an infinite number of positions available. If we define a small range of positions dx, then we can reasonably talk about the probability of landing in that range. The intensity is proportional to the square of the wave function, so we have:

$$\text{probability of particle landing in tiny range of positions from } x \text{ to } x+dx: \quad P\left(x \leftrightarrow x+dx\right) = \left[\psi\left(x\right)\right]^2 dx \tag{6.2.2}$$

We are taking one liberty here, in that we are ignoring the time dependence of this wave function. The point is that we are waiting "long enough" for the electrons to get to the screen after passing through the double slit, and interpreting the distribution of the dots on the screen as a probability distribution – we are not (yet) considering the time evolution of the wave function as it makes the trip.

Closer examination of the equation above reveals that it can't quite be right. The reason is that our wave function is *complex-valued*, and complex numbers do not necessarily have squares that are positive, and all probabilities must be positive! We therefore make a small adjustment to the "square the amplitude" prescription for intensity in this case: We take the square of the *magnitude*:

$$P\left(x \leftrightarrow x+dx\right) = \left|\psi\left(x\right)\right|^2 dx \tag{6.2.3}$$

The magnitude-squared of a complex number is the product of that number with its complex conjugate. The complex conjugate of a complex number is found by changing the sign of the imaginary part of the number:

$$Z = a+bi \quad \Rightarrow \quad Z^* = a-bi \quad \Rightarrow \quad \left|Z\right|^2 = Z^* Z = \left(a-bi\right)\left(a+bi\right) = a^2 - \left(bi\right)^2 = a^2+b^2 \tag{6.2.4}$$

The probability of the particle landing in a finite range (say between x_1 and x_2) is simply the sum of all the probabilities of landing in all the tiny ranges between those two points:

$$\text{probability of particle landing in range from } x_1 \text{ to } x_2 = P\left(x_1 \leftrightarrow x_2\right) = \int_{x_1}^{x_2} \psi^*\left(x\right)\psi\left(x\right)dx \tag{6.2.5}$$

The quantity $\left|\psi\left(x\right)\right|^2 = \psi^*\left(x\right)\psi\left(x\right)$ is called the *probability density*, for obvious reasons – integrating it over a range of positions gives the probability of the particle landing in that range.

Some Matter Wave Definitions

It's helpful to define some quantities that will help us manage the bookkeeping of rather complicated-looking wave functions (also so that we can understand what we read elsewhere!). There is no new physics here, just new language.

$$\text{angular frequency: } \omega \equiv \frac{2\pi}{T} = \frac{E}{\hbar} \tag{6.2.6}$$

$$\text{wave number: } k \equiv \frac{2\pi}{\lambda} = \frac{p}{\hbar} \tag{6.2.7}$$

These definitions make the expression of the wave function for a free particle (also called a *plane wave*, as it only moves in one direction, and regions of fixed phase form planes perpendicular to that direction):

$$\psi_{free}\left(x,t\right) = \psi_o e^{i\left(kx-\omega t\right)} \tag{6.2.8}$$

[*Note: By "free", we mean that it is not under the influence of any forces. We will later see how to deal with these situations – the wave equation (and therefore the wave functions that come from it) is different.*]

Alert

We have to be careful about associating matter waves too closely to "standard waves." For example, the speed of a standard wave is simply the wavelength divided by the period. This is known as the "phase velocity" of the matter wave, and it should not be confused with the speed of the particle (when it is observed as a particle). Indeed:

$$v_{phase} = \frac{\lambda}{T} = \frac{E}{p} = \frac{\frac{1}{2}mv_{particle}^2}{mv_{particle}} = \frac{1}{2}v_{particle}$$

For light, the phase and particle velocities do come out to be equal, but this is not so for massive particles. The way to think of it is this: The phase velocity is the rate at which the probability wave travels, but if we were to actually watch the particle (by reflecting light off it as it moves), it travels at a different rate. The phase velocity is not important in that it is not measurable – but be careful not to confuse these two quantities.

CHAPTER OVERVIEW

7: Quantum Mechanics in 1-Dimension

Index

G

Gedankenexperiment
2.2: The Nature of Time

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: 36

All licenses found:

- CC BY-SA 4.0: 72.2% (26 pages)
- Undeclared: 27.8% (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: Light as a Particle - *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*
  - 6: Matter as a Wave - *CC BY-SA 4.0*
    - 6.1: From Light to Electrons - *CC BY-SA 4.0*
    - 6.2: Interpreting Matter Waves - *CC BY-SA 4.0*
  - 7: Quantum Mechanics in 1-Dimension - *CC BY-SA 4.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*