Contents

1 Overview and Methods 3

2 Vectors and Statics 9

3 The Analysis of Motion 15

4 Force and Equilibrium 21

5 Newton's Laws of Motion 27

6 Circular Motion and Gravity 33

7 Energy and Action 37

8 Momentum and Collisions 43

9 Rotation and Non-Inertial Frames 49

10 Torque and Free Rotation 55

11 Celestial Mechanics 63

12 Harmonic Motion 69

13 Elasticity 75

14 Fluids 81

15 Heat and Temperature 87

16 Kinetic Theory 93

17 Thermodynamics 99

18 Wave Motion and Radiation 105

19 Wave Motion and Interference 109

20 Geometric Optics 115

21 Physical Optics 121

22 Limits of Classical Mechanics 127

23 The Electric Field and Potential 131

24 Basic Electronics and Magnetism 137

25 EM Induction and AC Electronics 143

26 Electromagnetism and Relativity 149

27 Quantum Mechanics 157

28 Atoms and Solid State Physics 163

29 Nuclear Energy 169

30 High Energy Physics 175

Lecture 1

Overview and Methods

Read sections 1.1–1.3, also review Appendix

Frequently physics is characterized as the study of matter and energy. In more basic terms, we ask two fundamental questions:

- What are things made of?
- Why do they do what they do?

Here is the short answer.

- Everything is made of elementary particles (quarks and leptons) subject to four fundamental interactions (electromagnetism, gravity, and two nuclear forces).
- Other than gravity, these three interactions are mediated by the exchange of particles called "bosons". The electromagnetic boson is the photon which is a particle of light.
- Gravity alters the structure of space and time causing all matter to coalesce and, unless opposed, ultimately collapse. It is unclear whether gravity can be modeled with an exchange boson like the other three interactions.
- These interactions cause the material particles to combine in certain ways (nuclei, atoms, molecules) and those combinations in turn interact in secondary ways subject to the science of mechanics.
- On a small scale we use quantum mechanics, when things get very fast or energetic we use relativity, otherwise we can use classical mechanics summarized in Newton's three laws of motion.

In physics we focus on measuring simple things. This is one of the reason why physics is so successful: we only focus on the very simplest of systems. No complications of living things, no historical accidents and lacuna to deal with, perfect repeatability.

Given this self-imposed restriction, it's not too surprising that physics is so accurate. What is surprising is the scope to which this accuracy extends. Today, this accuracy extends to all known physical experiments. In other words, there is no known experiment that cannot be explained by the current theories. There are areas for improvement, but the theories are predictive in every case up to experimentally measured limits. Einstein once said,

> The most incomprehensible thing about the world is that it is at all comprehensible.

One of the truly amazing things about physics is the reduction of such a huge variety of phenomena to these basic laws of mechanics. The story line is roughly aligned with the year-long structure of this class.

In the first term, we learn the principles of Newtonian mechanics including the ideas of energy, momentum and rotation. This term is quite linear: each lecture builds on the previous. We start with analyzing rudimentary things like rocks in free-fall or balls rolling down slopes.

The second term focuses on the basic properties of matter and energy. Whether it's solids, liquids and gases, or heat, sound and light, each branch has a unique way of approach in to the subject. We start by learning each approach then we discover how these rules follow from Newton's laws. Each of these branches has its own story, so this term is less linear than the first.

The third term involves the study of electricity and magnetism which we will find underlies all the mechanical forces (except gravity and weight). We will find we need to correct Newton's laws with those of Einstein's relativity. After this, we end with the solution to a serious problem: atoms that obey Newton's (or Einstein's) laws of motion and electromagnetic theory cannot exist! Quantum mechanics solves this problem and involves a more accurate (though counter-intuitive) understanding of the subatomic world.

So, we begin and end with mechanics. The goal is to explain the stuff in the middle with these laws of motion.

In order to quantify motion, we need to talk about how to precisely measure distance and duration. These precision measurements are what gives us the data to refine, verify, and use the laws of motion.[1]

Realize that ultimately every measurement in physics is a measurement of either distance or time. Whether it is the measurement of how far an object falls, or how far a needle moves on some meter, it is a distance. The measurement of time usually involves the comparison to some cycle: the earth around the sun, the vibration of quartz, etc.

Now measurement involves a comparison to some conventional unit. In other words, there are two things to consider: the unit and the method of comparison. A well-chosen unit will ease the method of comparison by making the process universal, portable, and stable. This is why over the years, even though the metric system of units doesn't change, sometimes the definition of these processes will. For example, the unit of length (the **meter**) used to be defined via a platinum-iridium bar held in Paris,[2] but in 1960 this definition was replaced with 1,650,763.73 wavelengths of the orange-red emission line of the krypton-86 atom in a vacuum. In 1983 this definition was replaced with the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second. This fixes the definition of the meter to the definition of our unit of time. Originally, the **second** was defined as 1/86400 of a solar day. Now the definition of the second is 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

These considerations are irrelevant for us since our experiments and the problems we will discuss require nowhere near this kind of precision. A good old stop watch and meter stick will be sufficient. But all of this hoopla is a testament to the precision and accuracy of modern-day physical science. What is relevant to us is learning how to use the metric system: conversion from other systems of units (e.g, inches and miles) and scientific notation.

I can never quite remember how to do these conversion calculations. I always remember one basic idea: each conversion equation is equivalent to a fraction

[1] It is also true that hidden assumptions about these techniques of measurement are why we need to replace Newton's laws with relativity and quantum mechanics in the third term.

[2] In 1793, this bar was intended to be one ten-millionth of a quadrant of the earth's circumference. It was later determined that this bar was short by a fifth of a millimeter. This makes the point: which is the unit, the bar or the earth? The earth is universal, but the bar is easier to use. So the bar is the true standard even if it is built incorrectly. See here for more details.

equal to one. In other words, we have

$$1 \text{ m} = 39.37 \text{ in} \iff 1 = \frac{1 \text{ m}}{39.37 \text{ in}}$$

Suppose I have something that is 12 inches long. The way to determine its length in meters is to "multiply by one":

$$12 \text{ in} = 12 \text{ in} \times \frac{1 \text{ m}}{39.37 \text{ in}} = 0.3048 \text{ m}$$

Notice how the units in the fraction are designed to "cancel out". So the process can work the other way too. What is the length of 0.8 meters in inches?

$$0.8 \text{ m} = 0.8 \text{ m} \times \frac{39.37 \text{ in}}{1 \text{ m}} = 31.50 \text{ in}$$

Here is a more complicated example. Convert 25 miles per hour into meters per second. We need to look up the conversion from miles to meters (or vice versa). I find 1 mile = 1609 meters. Since there are 3600 seconds in a hour, the calculation is:

$$\begin{aligned}25 \text{ mph} &= 25 \ \frac{\text{mi}}{\text{hr}} \times \frac{1609 \text{ m}}{1 \text{ mi}} \times \frac{1 \text{ hr}}{3600 \text{ s}}\\ &= 11.17 \text{ m/s}\end{aligned}$$

One mistake that people often make is the conversion of squared and cubed units. No matter how many times I explain it, someone always makes this mistake. But the approach is no different than the above. Suppose I want to know how many square centimeters are in one square inch. I have:

$$\begin{aligned}1 \text{ in}^2 &= 1 \text{ in}^2 \times \frac{2.54 \text{ cm}}{1 \text{ in}} \times \frac{2.54 \text{ cm}}{1 \text{ in}}\\ &= 6.452 \text{ cm}^2\end{aligned}$$

Notice how there are two conversion factors since a square inch is really an inch multiplied by an inch. For a cubic unit there are three conversion factors. For example: there are one million cubic centimeters in a cubic meter.

We also need a way of dealing with very large and very small numbers. This is necessary when one speaks of the number atoms in a glass of water, the distance to far-off galaxies, the size of the wavelength of light, etc. There are two approaches: use scientific notation or use metric prefixes. In reality you will need to do both, although technically one could get by with only one approach. Scientific notation is a bit cumbersome to get used to, but it is also a bit easier to use once you get used to it, so this is really the preferred method—especially for very large or small numbers. A number in scientific notation looks like this:

$$1.23 \times 10^{-45}$$

Sometimes this is also written as 1.23E-45 which is much easier to type and is usually how calculators display the number. I won't belabor this topic—I think you've all seen scientific notation before.

The other approach is to use the metric prefixes. This is much easier to talk about and sometimes easier to visualize. The basic idea is to attach a prefix to the unit which represents a multiple of 1000. So a kilometer is 1000 meters, a millisecond is 1/1000 of a second. The most commonly used prefixes are in Table 1.1.[3]

Sometimes "u" is used as a simple text replacement for the μ in micro. In addition, there are also prefixes between 10^{-3} and 10^{3}, but the only time you will ever see them is in the centimeter, which is 10 millimeters. For example, a cubic centimeter is a convenient unit of volume.

[3] See here for a complete list.

Prefix	Symbol	10^n	Example
tera	T	12	terabyte: size of large computer hard drive
giga	G	9	gigawatt: nearly enough energy to operate a flux-capacitor
mega	M	6	megahertz: frequency of radio waves
kilo	k	3	kilometer: largest commonly used length
milli	m	-3	millimeter: smallest commonly used length
micro	μ	-6	micrometer (a.k.a. micron): size of transistor
nano	n	-9	nanometer: size of the atom
pico	p	-12	picosecond: speed of computer calculations
femto	f	-15	femtometer: size of the nucleus

Table 1.1: Commonly used metric prefixes

Like any science, physics is a combination of deductive and inductive elements. Deduction works from evident principles to particular predictions. An example is Newton's laws of motion. Start with three laws and deduce the future motion of a particular situation. Deduction builds a hierarchy of laws and theorems each applicable to various branches of the science. So, hydrodynamics is a special case of Newton's laws applied to fluids.

Induction works the other way. From particular examples we tease out a pattern from the results. This almost always takes the form that when x changes so does y. If both are quantified, this relationship is summarized by the function $y = f(x)$. If the driver variable x occurs in a small range[4], this function can be written as $y = kx$. The value of k is called a **proportionality constant**. Frequently the first inductive step is to identify the x and y and use the data to calculate k.

Sometimes these processes are divided into a kind of division of labor: theorists work on the deductive side, experimentalist work on the inductive side. Of course, the truth of the matter is not so cleanly divided. But roughly we can say that the experimentalists calculate the values of k and the theorists try to derive the value of k from first principles. The extent to which these two groups agree represents the success of the science.

The purpose of many of the labs in this class is to perform just this comparison. It is the nature of induction that no one single experience can prove anything.[5] The only thing we can do is to build a preponderance of evidence. We run the experiment again and again, controlling as many variables as we can to isolate the x and y in which we are interested. If the data lines up, we are happy. But, of course, the data never does. In fact, each experiment is subject to some sort of unavoidable measurement error, so there really is no "line". It is the role of statistics to tease out the pattern in the data and estimate the size of the these errors in this analysis.

One thing that intimidates many people who have never been exposed to physics class is the math. Most of the homework and exams involve word problems which are notoriously difficult for students. But this is unavoidable. Physics without math is like music without notes, like football without the ball. Mathematics is the very language of physics. Going back to the ancient Greeks, the highest level of mathematics has always been brought to bear on physical questions. In some cases physical inquiry has driven mathematical development. To truly work in modern physics one must know about group theory, differential manifolds, and a host of other exotic mathematical concepts. In fact, it's not unreasonable to say

[4] How small? Small enough to make this statement true.

[5] But a single experiment can falsify it. This insight is often attributed to Karl Popper. However, not just any experiment will do. You will often find in lab that your experiments do not match the textbook theory. This is always a result of your lack of skill as an experimentalist. Sorry to be so blunt—one of the purposes of these labs is to develop these skills in you. Sometimes you'll just run out of time in a particular lab to get it right.

that the limit of your math ability will be the limit of your physics ability. I don't think it is any coincidence that the rise of modern physics in the 16th century (usually Galileo Galilei is chosen as a starting point) occurs about the same time as the rise of modern algebra (for example, see François Viéte).

But here is the silver lining: these highly refined levels of math are only required to get to the very edge between what we know and what we don't. It is very possible to understand the central core of the science with basic algebra, geometry and a dollop of calculus.

You'll need to know how to solve basic algebraic equations—we will be doing this all the time. Occasionally we will need to solve a system of equations: two equations with two unknowns. Occasionally we will need to solve an equation involving logarithms.[6] That's about the extent of the algebra. I'll assume you can do the basics and will table reviewing the "occasionals" until we need them.

You will also need to remember some geometry (or at least how to look them up). Circles, triangles, parallel lines. Things like $C = 2\pi r$, $A = \pi r^2$, the angles of a triangle add to 180°, the Pythagorean theorem, etc. One nice trick to remember is that when a third line is drawn across a pair of parallel lines, the opposite angles are equal—both on the inside of the parallel lines and across each intersection (see the left hand side of Figure 1.1). Another fact we will use frequently is that the angle between two lines is the same as the angle between their perpendiculars (see the right hand side of Figure 1.1). We will also use a lot of trig, but we will review all that in Lecture 2.

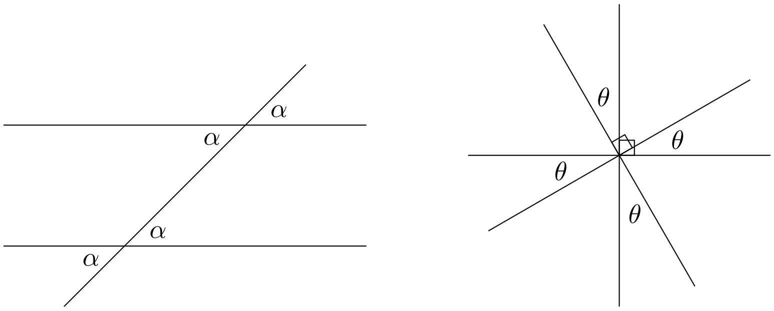


Figure 1.1: Equal angles with parallel lines and equal angles between perpendiculars

As far as calculus goes—really we should know some. Don't worry: I know this is a non-calculus class. But physics is the study of change and calculus is math designed to describe change. So, I'll mention a couple of things. As I said before, any causal relationship between measurable quantities can be expressed as a mathematical function, $y = f(x)$. Frequently we will be interested in the following question. If I change x a small amount (call it Δx), how much does y change (call it Δy)? In general, we have $\Delta y = k(x)\Delta x$. Now if I rewrite this as

$$k(x) = \frac{\Delta y}{\Delta x} \tag{1.1}$$

we call $k(x)$ the **derivative** of $f(x)$. If I were to plot $f(x)$ on a graph, the value of $k(x)$ corresponds to the slope of the curve at that point (see Figure 1.2). So, the derivative represents how sensitive $f(x)$ is to changes in x.

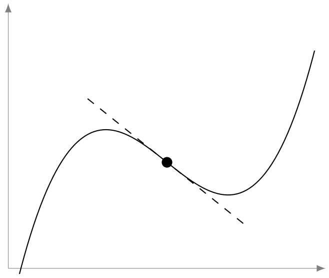


Figure 1.2: Definition of derivative

The ways in which calculus are used in physics are wide ranging (you could argue that calculus was invented for the sake of solving physics problems). But one trick is particularly helpful. If you know the value of a function at a particular point and the derivative at that point, the values in that neighborhood is given by

$$f(x + \Delta x) = f(x) + k(x)\Delta x \tag{1.2}$$

This is really just rewriting the definition of the derivative. But how do we calculate these derivatives? Well, you need to take a calculus class to know that.

[6] Exponential functions are used to describe processes that involve growth or decay. Logarithms are typically used to solve these kind of equations.

But I can tell you one thing. The derivative of x^n equals nx^{n-1}. So, here's a trick: what's the square root of 40? In symbols:

$$y = \sqrt{40} = (36 + 4)^{0.5}$$

Now, the derivative of $x^{0.5}$ is $0.5x^{-0.5} = 0.5/\sqrt{x}$, so

$$y = \sqrt{36} + \frac{0.5}{\sqrt{36}} \times 4 = 6\tfrac{1}{3}$$

The real answer is 6.3246—an error of less than 0.2%.

That's enough for now. I'll let you know some more calculus tricks as we go along. But be assured: you won't be expected to really know these details for this class. Consider them the special spices in the luscious banquet that is this class.

One particularly important application of these ideas is in a formula called the **binomial theorem**. Using the idea of a derivative it is possible to show that

$$(1 + x)^n = 1 + nx^{n-1} \tag{1.3}$$

when x is very small. This formula is used in Lectures 13, 15, 23, and 26.

There is one last topic to touch upon: significant figures. These are just the number of digits we are willing to show in our final calculations. The point here is that every measurement involves some sort of uncertainty. When I use a meter stick to measure the length of an air track, is it reasonable to believe that I know this length down to the micron? No. Maybe down to the millimeter or so. If that's the case I better not write down that the length is 602.5 millimeters. The .5 is unjustifiable. In fact, the implied possible error associated with recording a measurement like 602 millimeters is ±0.5 millimeters. That is, I'm confident its neither 603 nor 601—but whether its 602.4 or 601.9, I simply don't know.

Now, if I use this number to calculate another—say I multiply by π—the result cannot be more accurate than the numbers going in! So even if your calculator says the answer is 1,891.2386 millimeters, this number conveys too much accuracy. We need to round the number to 1,890 millimeters. The basic rule is to round to the number of digits going in (in this case three). The zeros can sometimes make this confusing, but the basic rule is fairly straight-forward. If you can't figure any of this out, round to three digits—it'll be the right choice 80% of the time.[7]

Next week we will talk more about measuring distances and learn some new math called vectors. We will see that we can apply these ideas to objects in equilibrium under several forces. This will give us the chance to talk about force, mass and weight.

[7] However, you should keep more digits when performing intermediate calculations. Rounding can introduce an error that propagates through the calculation. In general, keep five digits until the very end. Then round to three.

Lecture 2

Vectors and Statics

Read sections 1.4–1.8, sneak a peak at sections 4.11, 7.4, 9.2, 18.5, and 21.2

We live in a three-dimensional world. So, as we start to describe how things move in space we need to take into account both distance and direction. Since geometry studies the properties of space, it's natural to expect the language of physics to be geometric. If you pull a copy of Newton's Principia from the library or Internet, you will see that it is dominated by classical geometric theorems and reasoning. Fortunately, we don't need to know that much geometry. The invention of vectors and vector notation (usually associated with Josiah Gibbs) greatly simplifies the reasoning required to solve physics problems.

The reason why vectors are so much easier is that they convert geometric problems into algebra. The archetypal example of a vector is a simple displacement from here to there. This is usually drawn as a little arrow from point A to point B. The arrow is important because we want to maintain a distinction between the displacement that goes from A to B and that which goes from B to A. Typographically some authors use bold letters to represent a vector, but I prefer to use a letter with an arrow over it, like $\vec{a}$, which is pretty easy to write.

Now consider a two-fold movement from point A to point B then to point C. We can represent this motion as two vectors in space, $\vec{a}$ pointing from A to B and $\vec{b}$ pointing from B to C. The whole motion can be captured in a vector pointing from A to C—we'll call it $\vec{c}$. See Figure 2.1. When three vectors are associated in this way we say that $\vec{a}$ and $\vec{b}$ add up to $\vec{c}$. In symbols we write

$$\vec{c} = \vec{a} + \vec{b}$$

which looks just like adding numbers. The suggestion is deliberate, but remember: vectors are not numbers! Every vector equation like this has behind it a triangle like Figure 2.1.

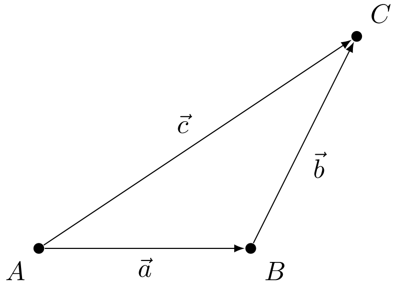


Figure 2.1: Vector addition

The reason we call this **vector addition** is that this way of combining vectors obeys the laws of arithmetic. For example, it commutes:

$$\vec{a} + \vec{b} = \vec{b} + \vec{a}$$

In order show this, I need to clarify one thing. The essence of the vector is its distance and direction, not where it sits. In order to represent this equation we need to move the arrows so that the tail of the second is on top of the head of the first. In other words, draw the first vector with the proper length in the correct direction then draw the second vector in the same way. If we do this with the vectors we were using previously we would get a diagram something like Figure 2.2. Notice how the pairs both end up in the same spot. You can see from this figure why vector addition is sometimes said to obey the **parallelogram law**.

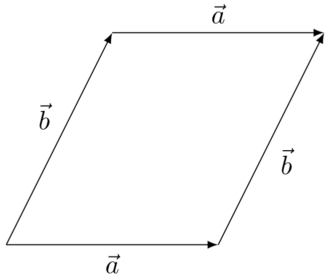


Figure 2.2: Vector addition commutes

Vector addition is also associative, has a zero and inverses. The zero vector has no length and no direction. The inverse of a vector is the vector that points in the exact opposite direction, so if $\vec{b}$ is the inverse of $\vec{a}$, we have $\vec{a} + \vec{b} = 0$.[1]

There is also another thing we can do with vectors called **scalar multiplication**. Suppose I take a vector $\vec{a}$ and add it to itself. I get another vector in the same direction, with twice the length. In fact, I can write

$$\vec{a} + \vec{a} = 2\vec{a}$$

The two is kind of "multiplied" into the vector just like in basic arithmetic: $5 + 5 = 2 \times 5$. Geometrically, scalar multiplication stretches (or shrinks) the size of the arrow, but algebraically it acts like multiplication and obeys the standard laws of arithmetic. In particular, negative one will flip the direction of the arrow making it point in the opposite direction. This even allows us to subtract vectors. Refer back to Figure 2.1. I can write the following vector subtraction from this diagram: $\vec{c} - \vec{b} = \vec{a}$. The equation says: run up $\vec{c}$ then move backward on $\vec{b}$ and you end up where $\vec{a}$ ends up. I think you can start to see how these little arrows are forming a true algebra.

It's useful to have this vector representation for understanding the basic concepts of physics. However, there is one more step we need to take to unleash their full power. We need to talk about **vector components**. This approach is similar to the use of Cartesian coordinates to describe where a point is in the plane. Each Cartesian grid defines a couple of unique vectors. Consider a vector that points in the x-direction with a length of one unit. Usually this vector is denoted $\hat{x}$ (pronounced "x-hat"). The little caret on top indicates that this is a unit vector—a vector with the length of one. Similarly there is the unit vector that points in the y-direction denoted $\hat{y}$.

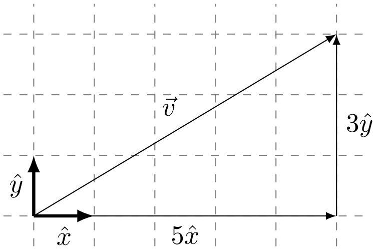


Figure 2.3: Vector basis in action

These two vectors are called a **vector basis** for the plane because they are sufficient to describe any other vector in the plane. For example consider the vector $\vec{v} = 5\hat{x} + 3\hat{y}$ (see Figure 2.3). We say that its x-component is five and its y-component is three. These quantities are usually denoted v_x and v_y respectively.

Every vector that can be drawn in the plane can be completely specified by these components. In almost every problem, we will be utilizing the components of vectors to perform our necessary calculations. The reason for this is that every vector equation implies an equality between components. In other words,

$$\vec{a} = \vec{b} \implies a_x = b_x \quad \text{and} \quad a_y = b_y$$

In general, each vector equation creates a component equation for each dimension of the problem.

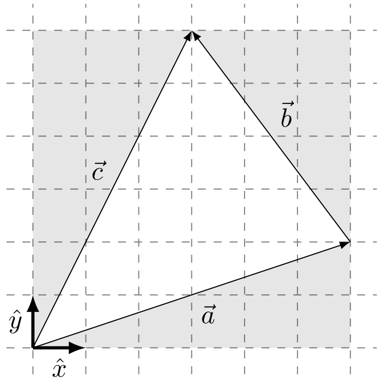


Figure 2.4: Vector addition using components

This makes adding two vectors easy once I know their components. Let's take $\vec{a} = 6\hat{x} + 2\hat{y}$ and $\vec{b} = -3\hat{x} + 4\hat{y}$ and call their sum $\vec{c}$:

$$\vec{c} = \vec{a} + \vec{b}$$

Using components, it's easy to see what $\vec{c}$ is:

$$\begin{aligned} c_x &= a_x + b_x = 6 + (-3) = 3 \\ c_y &= a_y + b_y = 2 + 4 = 6 \end{aligned}$$

So, $\vec{c} = 3\hat{x} + 6\hat{y}$. You can draw the triangle to double-check. You should get something like Figure 2.4.

What is the length of this vector $\vec{c}$? This also is easy to answer now that we know its components. Look again at Figure 2.4. Notice how the shaded triangle next to the vector $\vec{c}$ is a right triangle? In fact, the hypotenuse of this triangle is the length we are interested in—and we know the length of the sides because they are the components we just calculated.

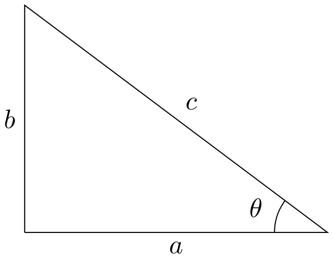
[1] I am too lazy to write the zero vector as $\vec{0}$ even though I probably should.

Using the Pythagorean theorem, the length[2] is

$$c = \sqrt{c_x^2 + c_y^2} = \sqrt{(3)^2 + (6)^2} = 6.7$$

What about its direction? Now we need to talk trig...

Remember for any right triangle, the three basic trig functions relate the sides of the triangle to the angle inside the triangle. By definition, the sine of an angle is the ratio of the side adjacent to the angle and the hypotenuse (which is opposite to the right angle). The cosine is the ratio of the opposite side and the hypotenuse. The tangent is the ratio of the opposite and the adjacent (which is also the ratio of the sine and cosine). These relations are summarized in the familiar Figure 2.5.



$$\sin\theta = a/c$$
$$\cos\theta = b/c$$
$$\tan\theta = a/b$$

Figure 2.5: Trig function definitions

So, if we want to know the direction of our vector $\vec{c}$ in Figure 2.4, we can use the tangent function. The components are the adjacent and opposite sides, so we have:

$$\theta = \tan^{-1}(c_y/c_x) = \tan^{-1}(6/3) = 63.4°$$

We can also turn things around. If we only know the direction and length of a vector $\vec{v}$, the components are given by the other trig functions:

$$v_x = v\cos\theta$$
$$v_y = v\sin\theta$$

You will use these equations again and again in this class, so make sure you understand them.

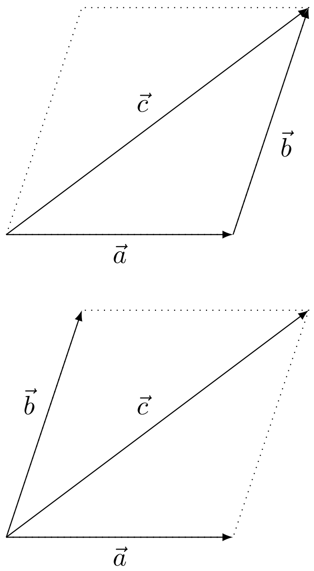


Figure 2.6: Consecutive versus simultaneous vector addition

Until now I have been using displacements in the plane as my example of vectors. These are not the only physical quantities that can be represented by vectors. Anything involving a direction can typically be represented by a vector. One quantity of particular importance is force. As I mentioned in the Lecture 1, the idea of force runs through all of classical mechanics; Newton's laws are built to understand the effects of various forces. All that I have said up to now also can be applied to the forces operating on an object with one subtle distinction.

In describing vector addition with displacements I asked you to imagine the two vectors being combined as head-to-tail, or consecutive. With force it is more appropriate to imagine the vectors combining simultaneously. In the end, this distinction is not important in our calculations, but it does change how we draw the combinations. See Figure 2.6 for what I mean.

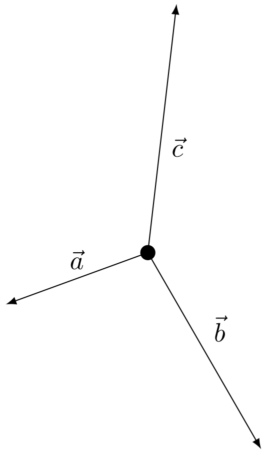


Figure 2.7: An object pulled by three forces

A typical force problem is like this: An object is pulled in three directions by three forces (see Figure 2.7). The force $\vec{a}$ has a magnitude of 2 at an angle of 200° and the force $\vec{b}$ has a magnitude of 3 at an angle of 300°. What must the magnitude and angle of the third force be to balance the other two?

In vector notation, the answer is simple. We want all three forces to balance—in other words, we want them all to add to zero:

$$\vec{a} + \vec{b} + \vec{c} = 0$$

[2] The length of a vector is commonly denoted by the same letter without the little arrow on top.

We solve for $\vec{c}$ and we are done:

$$\vec{c} = -(\vec{a} + \vec{b})$$

But not really. We want to know both the magnitude and direction of $\vec{c}$. We will get those by calculating components. The components of $\vec{a}$ are:

$$\begin{aligned} a_x &= a\cos\theta = (2)(\cos 200^\circ) = -1.88 \\ a_y &= a\sin\theta = (2)(\sin 200^\circ) = -0.684 \end{aligned}$$

Notice how they are both negative. This is as it should be because the positive x-direction is to the right but this vector points to the left. Similarly, the positive y-direction is up but this vector points down. It's always a good quick double-check to make sure these signs are right. The components of $\vec{b}$ are:

$$\begin{aligned} b_x &= b\cos\theta = (3)(\cos 300^\circ) = 1.50 \\ b_y &= b\sin\theta = (3)(\sin 300^\circ) = -2.60 \end{aligned}$$

So, the components of $\vec{c}$ are:

$$\begin{aligned} c_x &= -(-1.88 + 1.50) = 0.38 \\ c_y &= -(-0.684 + -2.60) = -3.28 \end{aligned}$$

Therefore its magnitude and direction are

$$\begin{aligned} c &= \sqrt{(0.38)^2 + (-3.28)^2} = 3.31 \\ \theta &= \tan^{-1}(-3.28/0.38) = 83.4^\circ \end{aligned}$$

Those are the basics on vectors. As you can see, most of the time you will be calculating the components of given vectors, adding those components, and occasionally converting these answers back into a magnitude and direction. The hardest part is keeping the trig straight.

But I find it hard to stop here without covering a few supplemental topics...

There are other ways to combine vectors. These are also called multiplication, but I think that this is just a way to distinguish them from the more fundamental operation of vector addition.[3] The first is called the **dot product** and it has two equivalent definitions:

$$\vec{a} \cdot \vec{b} = ab\cos\theta$$

where θ is the angle between the two vectors. This is the "geometric" definition in terms of lengths and angles. The "algebraic" definition is in terms of components:

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y$$

It's not obvious that these two definitions are equivalent but they are. This vector combination is useful when talking about work and energy in Lecture 7.

A second combination is called the **cross product** (a.k.a. the **vector product**). The "geometric" definition is:

$$\vec{a} \times \vec{b} = (ab\sin\theta)\hat{n}$$

where $\hat{n}$ is the unit vector that points perpendicular to the plane defined by $\vec{a}$ and $\vec{b}$. The "algebraic" definition is

$$\vec{a} \times \vec{b} = (a_x b_y - a_y b_x)\hat{n}$$

(This assumes that $\vec{a}$ and $\vec{b}$ lie in the xy-plane, $\hat{n}$ points out of the page). Notice that the dot product produces a number but the cross product produces a vector.

[3] Although they do distribute over vector addition, so this nomenclature is not without merit.

This vector combination is useful when talking about torque and rotation in Lecture 10 and also magnetism in Lecture 24.

A few introductory physics texts mention these two vector products, but none talk about tensors. I'm not sure why since they aren't too hard to understand. A **tensor** is a linear function between vectors. Remember, a function represents a causal relationship between variables. If the variables are represented by vectors, you may have a tensor on your hands. For this relationship to be a tensor it must be linear in the sense that it must preserve both vector addition and scalar multiplication. In symbols, a vector function f is a tensor when the following are true:

$$\begin{aligned} f(\vec{v}+\vec{u}) &= f(\vec{v})+f(\vec{u}) \\ f(a\vec{v}) &= af(\vec{v}) \end{aligned}$$

It can be shown that the basis in the underlying vector space also allows us to define tensor components to describe these functions. In our case there would be four (or nine if we are talking 3D). Tensors can be helpful in discussing rotation, elastic stress and electromagnetism.

Finally, it's worth mentioning **four-vectors**. One particularly concise way to deal with the complications of relativity is to use the idea of a four-dimensional vector. This is related to the fact that Einstein showed that it is not appropriate to consider space and time as separate entities but rather as a combined space-time continuum. This combination of the three dimensions of space with the fourth dimension of time propagates through the various concepts of physics in a way that is both elegant and surprising.

Next week we will develop some fundamental concepts that will allow us to describe motion in a way that will fit in with Newton's three laws of motion (Lecture 5). In particular we will find out the path of a projectile under the influence of gravity.

Lecture 3

The Analysis of Motion

Read sections 2.1–2.7 and sections 3.1–3.3

Any physical system is specified mechanically by its configuration: the relative position and orientation of its parts. The motion of the system is defined when its configuration is specified over time. What we need is a way to record this configuration. This is usually very difficult[1] and requires great cleverness on the part of the experimentalist. But some systems are simpler to characterize than others. The simplest of all is the one whose internal configuration is negligible. This system is called a **particle** and is defined by its position in space. A contraption that will record this position in space is called a **reference frame**.[2]

Suppose we set up a reference frame and begin to record the position of our particle. Obviously the values we get will depend upon the details of the reference frame—we will assume this frame is perpendicular, uniform and stationary. For now let's restrict ourselves to motion in a plane. Then the position is specified by the two numbers associated with the Cartesian grid in that frame. As we monitor the motion of the particle, these two numbers change. In other words, they are functions of time: $x(t)$ and $y(t)$.

Now consider the displacement of the particle between two moments of time. Let's agree to call the first instant t_0 and the second simply t. The position of the particle at each of these moments in time define a displacement vector which we will call $\Delta\vec{x}$. The x-component is given by $x(t) - x(t_0)$ or in more compact notation, $x - x_0$. Similarly for the y-component.

Notice that the length of the displacement $\Delta\vec{x}$ represents the net motion of the particle. The overall path-length traversed cannot be smaller and may be much longer than the net displacement. We will see that both the path-length and the net displacement are important quantities to track, but the displacement is more important. Above all, we must remember the distinction between the two!

The rate at which the position changes is called **velocity**. Velocity is a vector, so it has both direction and magnitude. The faster the speed of the object, the larger the magnitude of the velocity. The rate of change of any quantity is the ratio of the size of the change to its duration, so velocity is simply the displacement $\Delta\vec{x}$ divided by $\Delta t = t - t_0$.[3]

What I have just described is called the **average velocity**. Because the displacement $\Delta\vec{x}$ represents the net motion of the object, this velocity is a kind of average of the speed during the motion of the object. For example, if the object moves in a complete circle, the net displacement will be zero and so will the average velocity. This is because the velocity takes into account both the speed and direction. The velocity on the upper half of the circle is canceled out by the lower half.

[1] And frequently impossible when we talk about atomic theory.

[2] I am trying to emphasize the fact that this reference frame is not a mental construction but a physical one. We use the frame to construct a mental inventory of our system, but the raw data is coming from rulers and clocks subject to the laws of physics. These statements will become important when we talk about the need for modifying Newton's laws with Einstein's relativity in Lecture 26.

[3] Technically this is not a division, but a scalar multiplication of $\Delta\vec{x}$ by $1/\Delta t$.

More frequently we are interested in the velocity of an object at a particular moment in time. The problem we now face is that our definition of velocity requires us to consider two moments in time.[4] Clearly if we shrink the interval in time between these two moments, we get closer to what we want. This is where the calculus comes in. We want to "take the limit" as this interval Δt goes to zero. The challenge is that as the denominator of the velocity shrinks, so does the numerator—the displacement involved becomes smaller and smaller—we are going to end up with the value 0/0. Well, the purpose of calculus is to make sense of this nonsense.

But we don't need to know the technical details involved. Take it for granted that this process is well-defined. We call the velocity defined in this way the **instantaneous velocity**. As I mentioned above, this is what we will typically be interested in, so when you see the word velocity you should assume it refers to the instantaneous velocity of the object.

If we take the position of a particle and plot it against time we get a curve. We can say that the position is a function of time. Then the velocity is its derivative as defined in the Lecture 1. The velocity is the slope of the line at a particular point in time (see Figure 3.1).

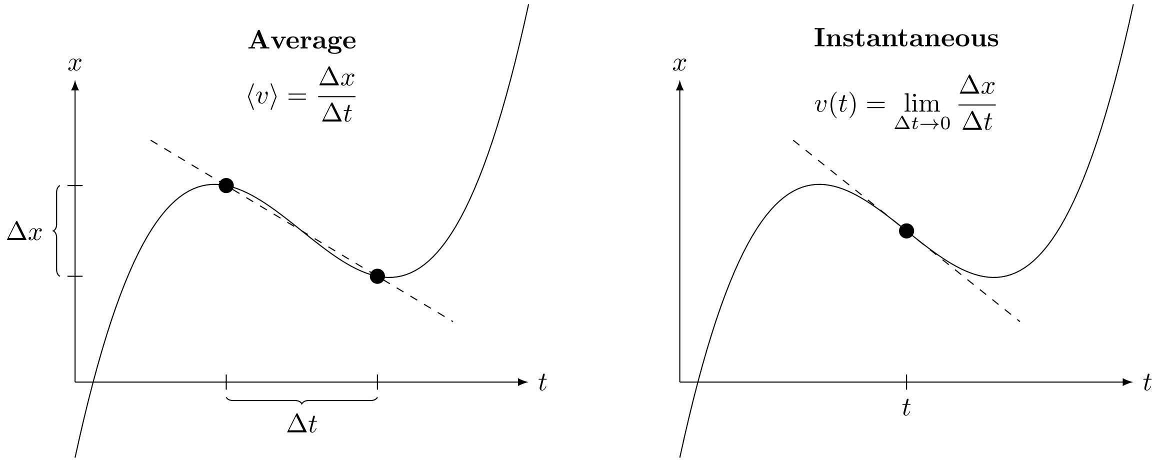


Figure 3.1: Velocity defined as the slope of the space-time curve

Frequently in these motion-based physics problems we will be given the initial velocity of an object and our goal will be to describe some characteristic of the motion. The velocity tells us the direction in which the object will move and its speed. If it is moving at 30 m/s, we know that in one second it will have moved 30 meters. After two seconds it will have moved 60 meters. The formula for this is

$$x = vt \tag{3.1}$$

I like to call equation (3.1) the "constant velocity equation". You can see it is really a rearranging of the definition of velocity with x_0 and t_0 set to zero. Remember this formula only works if the velocity does not change. If either the speed or direction of the object changes during its motion, equation (3.1) will not work.

If the velocity does change, its rate of change is called **acceleration**. You might be less familiar with the notion of acceleration, but this quantity is the most important one for understanding motion. This is not immediately obvious and this insight is one of the keys to understanding Newton's laws of motion. We will see that a constant force will generate a corresponding constant acceleration in an object.

An important special case of this is an object's weight. This is the gravitational force the earth exerts trying to pull an object to its center. Newton discovered

[4] So does the lab. In the lab, velocity measurements always involve measuring the length traveled in a specific period of time and dividing the two.

the laws that govern gravitation and was able to explain the workings of the solar system with it—we will talk about that in Lectures 4, 5, and 10. But on the surface of the earth the force of gravity is fairly constant.[5] Because of this, the weight of an object causes it to fall with a constant acceleration of 9.8 meters per second squared, symbolized by g.

So it is worth spending some time on understanding the motion of particles with constant acceleration. Since the definition of acceleration parallels the definition of velocity, we must have a formula that parallels equation (3.1). It is

$$v = v_0 + at \tag{3.2}$$

Remember that this formula depends upon the acceleration being constant—which is not always the case. We will see an example of that in Lecture 6 with uniform circular motion. In that case we need a different formula.

But how do we relate equation (3.2) to position? Calculus is designed to deal with this type of problem—how do we calculate the position if the velocity is changing? However, there is another trick we can use because the rate at which the velocity is changing is constant. If a quantity increases (or decreases) at a constant rate, its average value is the average of the initial and final values.[6] So,

$$\langle v \rangle = \tfrac{1}{2}(v + v_0)$$

Now the average velocity by definition is the displacement divided by time, so we can rewrite this as

$$x = \tfrac{1}{2}(v + v_0)t \tag{3.3}$$

By combining equations (3.2) and (3.3) we can derive three more equations:

$$x = v_0 t + \tfrac{1}{2}at^2 \tag{3.4}$$

and

$$x = vt - \tfrac{1}{2}at^2 \tag{3.5}$$

and

$$v^2 = v_0^2 + 2ax \tag{3.6}$$

Equations (3.1)–(3.6) are the main results of this lecture. The rest of the lecture will be about applying them.

A typical problem is a object dropped from rest. Suppose we let a rock fall for two seconds. How far does it fall? In these constant acceleration problems there are five possible quantities to consider: t, x, v_0, v, and a. Notice how each of the equations (3.3)–(3.6) involve four of these five quantities. So one of the first steps in solving these problems is to identify the four quantities in the problem. For the rock problem we are told explicitly that the time involved is two seconds. We also know that the acceleration of the rock is -9.8 meters per second. Notice the negative sign. This is there to indicate that the acceleration due to gravity is down. What else? The initial velocity is zero. Occasionally you will need to tease this kind of implicit data from these problems. So we have three of four. The final quantity is the distance x—we don't know it, but we want to know it. The equation that involves t, a, v_0, and x is (3.4). We have:

$$x = (0)(2) + \tfrac{1}{2}(-9.8)(2)^2 = -19.6$$

In this case the negative sign is there because the net displacement is down. See Figure 3.2 for the space-time diagrams associated with this problem.[7]

So the general procedure is to identify the three pieces of data given (perhaps implicitly). Then look for the equation that involves those three and the one quantity you need. Then solve it.

Until now we have deliberately left out the idea of air drag. We have left it out not because it is negligible but because it is hard to deal with. Usually air drag

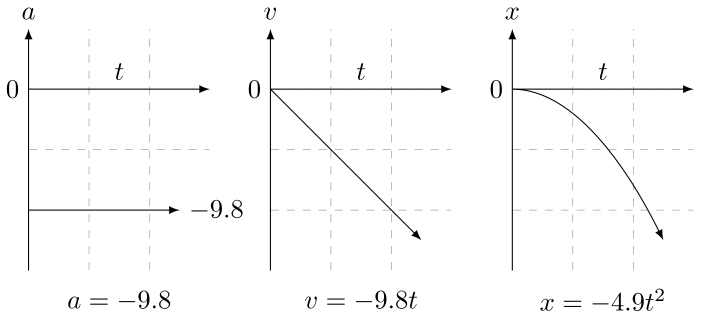
[5] Even in Newton's time it was possible to measure the slight variations in the force of gravity on the earth.

[6] I won't prove this statement, but consider the following example. Take the series 1,3,5,7,9,11. The average of these numbers is 6 which is also the average of the ends. Now take the series 2,4,8,16. The average of these numbers is $7\frac{1}{2}$ while the average of the ends is 9.

[7] Notice how when we have an object accelerated from rest, the distance it travels is given by equation (3.4) as

$$d = \tfrac{1}{2}at^2$$

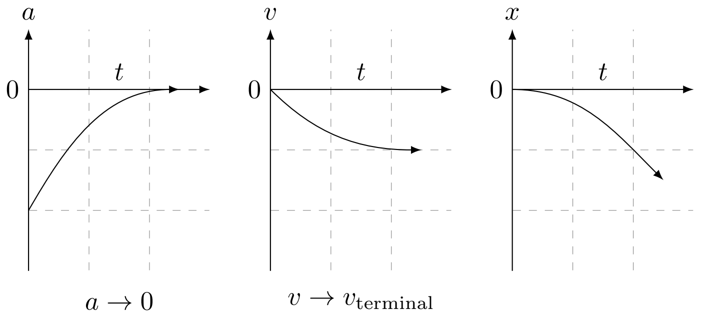
Figure 3.2: Space-time diagrams for free-fall with no air drag



introduces a deceleration[8] that is related to speed (the faster the speed, the larger the drag[9]). To solve these problems exactly requires some calculus. But we can get an approximate solution using a spreadsheet. The project for this term walks you through how to do this. In fact, the spreadsheet approach will work even for those problems when the calculus won't.

Although the detail of free-fall with air drag are complicated, its final state is easy enough to understand. Depending upon the details of the object and the air, a certain velocity will produce just enough drag to counter-balance the weight of the object. This is called the **terminal velocity** of the object. If the object starts with a velocity smaller than terminal (at rest, for example) then the net acceleration will increase the velocity. As the velocity increases, the drag will oppose more of the weight reducing the acceleration (slowing the rate at which velocity increases) until it reaches a steady-state at terminal velocity. See the space-time graphs in Figure 3.3 and compare them with those in Figure 3.2.

Figure 3.3: Space-time diagrams for free-fall with air drag



Until now we have only been discussing motion in one dimension. However, we can also discuss the motion of a projectile flying under the influence of gravity with these same equations. The problem of understanding the motion of projectiles goes back to antiquity and wasn't really solved until Galileo began to analyze the idea of acceleration. His main insight is that the vertical component of the motion is under the constant acceleration (due to gravity) while the horizontal motion has no acceleration—the horizontal motion obeys equation (3.1). This means that the trajectory in space is a parabola.

For example, suppose we launch a projectile at a 60° angle with an initial speed of 45 meters per second. How far will it fly? Ignore air resistance. See Figure 3.4 for reference.

The first step is to break the data into horizontal and vertical components. In the horizontal we know that we will use equation (3.1) which involves x, v_{0x}, and t. We are interested in solving for x and we are given enough information to solve for v_{0x}. This means we will be able to solve for t. This is common because the duration t is the same for the vertical and horizontal components. So the time is frequently a problem solving "bridge" between the information contained in the

[8] When the acceleration opposes the direction of motion (or velocity) we call it deceleration. Notice this is not the same as negative acceleration which indicates its direction in space. When an object falls, the deceleration from air drag points up.

[9] This is why you can't shoot fish in a barrel. The drag on the bullet is so great it can actually destroy the bullet itself. In fact, the higher caliber the more likely this will happen.

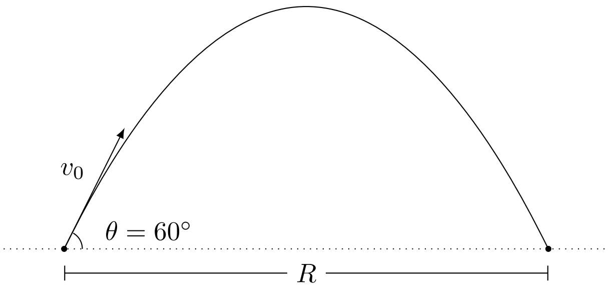


Figure 3.4: Calculating the range of a projectile

horizontal and vertical components. In our case we need to calculate v_{0x} first:

$$v_{0x} = v_0 \cos\theta = (45)(\cos 60°) = 22.500$$

But we can't calculate the time because we don't know the range. Since we have exhausted the information contained in the horizontal motion, let's turn to the vertical. We know v_{0y} is given by:

$$v_{0x} = v_0 \sin\theta = (45)(\sin 60°) = 38.971$$

Of course, $a = -9.8$ since this is a free-fall problem (no air drag). We want to determine t in order to use it in the horizontal calculation, so we need one more vertical quantity from the problem statement. The implicit data here is that $y = 0$ because we are asked about the range—the distance the projectile travels until it comes back to its original level. This is a net vertical displacement of zero. Given this information, we can use equation (3.4) to solve for t.

$$(0) = (38.971)(t) + \tfrac{1}{2}(-9.8)(t)^2$$
$$\implies t = 7.9533$$

Since the duration of the vertical motion is the same as the horizontal, we can finally solve for the total x-displacement. Thus,

$$x = (22.500)(7.9533) = 178.95$$

The original data was given with two significant digits, so we should round this final answer to 180 meters.

It is sometimes convenient to know the range of a projectile without going through the logic of the previous example. We can summarize the results in the formula[10]

$$R = (v_0^2/g)\sin 2\theta \tag{3.7}$$

This is called the **range equation**. Deriving formulas for every permutation of the projectile problem is tiresome, but this one can be useful on occasion.

There are a number of different of projectile questions that can be asked and sometimes finding the implicit data can get tricky. One trick in particular to note is that at the top of the trajectory the vertical velocity v_y is zero. Occasionally you'll need to use the quadratic equation. But the principles you will need to solve them are all laid out in this example.

Anyone who has played 30 seconds of golf will know that these problems are completely unrealistic and don't represent the true motion of the ball. The air not only resists the motion (air drag) but any spin on the ball will curve its path as well. The dynamics of spin and its curving effect are extremely complicated[11] but if we focus our discussion on just the air drag, we can say a couple of things.

First, the range of the ball is much shorter than without drag. The longer the time the ball is in the air the longer the effects of air resistance so the effect is most pronounced on projectiles with large initial velocities (there is also a

[10] You can derive this formula by simply using the same logic as our example problem but use letters without putting the numbers in. You'll need to remember the trig identity $\sin 2\theta = 2\sin\theta\cos\theta$.

[11] Ultimately it is based on Bernoulli's principle—something we will talk about in Lecture 14.

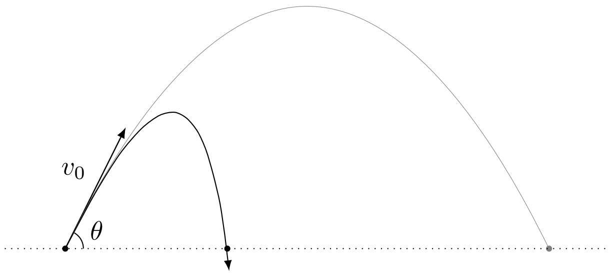


Figure 3.5: Projectile trajectory with heavy air drag

correspondingly larger air drag as well). It follows that the maximum height and the time of flight are also shorter. See Figure 3.5

Second, notice how the angle with which the ball falls is nearly straight down. This is because the direction of the air drag is in the opposite direction of the motion. In the horizontal we have a deceleration without any corresponding acceleration. So the air drag slows down the horizontal motion. Of course it does the same to the vertical, but the vertical motion has gravity to help out. What happens is that the projectile basically stops its lateral motion and falls downward at a steeper angle than it goes up. Anyone who has played an outfield position in baseball can attest to the fact that catching a ball hit out that far usually involves looking nearly straight up to catch it.

This is the real motion of projectiles and is the main reason it was so difficult to determine the true motion of a projectile. In the Middle Ages it was thought that a quantity called **impetus** was transferred to the object when it was thrown. After that, the motion was considered to be in a straight line as this impetus was used up (kind of like a wind-up toy). After the impetus was exhausted, the object would fall straight down following a trajectory like a triangle. I like to call this a "cartoon trajectory" because this is how things work in the cartoons: the bad guy runs straight off the cliff for a while then, all of a sudden (usually with a puff of smoke), he falls straight down to his doom. The moral here is that there is more truth to the cartoon trajectory than the clean parabolic motion the textbooks show.

Next week we will introduce the key concept of force into our vocabulary. Before we dive into the deep end of Newton's laws of motion we start by building some skills working with forces and vectors. We will learn the basic mechanical forces (weight, tension, support, and friction) and how they work. We will also learn what it means for these forces to be in equilibrium.

Lecture 4

Force and Equilibrium

Read sections 4.6–4.11, sneak a peak at sections 10.1, 18.5, 21.2, and review Lecture 2

When an object is subject to a force there are four physical quantities to distinguish: the net force, torque, stress and pressure. Technically, pressure can be seen as a special case of stress, but we'll set that consideration aside for now.

Force is the total amount of push or pull to which our object is subjected. The **pressure** is the force divided by the area across which it is applied. This is why the swami can lie of a bed of nails without getting hurt. The force of his weight is distributed, so that the pressure of support from any single nail is insufficient to pierce the skin. We won't need the concept of pressure again until Lectures 13 and 14. We will assume that any force is applied at a particular point on the object. Just recognize that this is an approximation: every real force is applied over a certain area.

Consider an object that is subject to a variety of forces all applied at different points. Each of these forces will have a tendency to do three things: (1) push the object in the direction of the force, (2) twist the object around its center of mass, and (3) deform the shape of the object.

If the direction of the force is precisely toward or away from the center of mass this force will not produce any kind of twisting or rotation. Otherwise the force is said to produce **torque** which will be discussed in detail in Lecture 10.

An object is said to be in **equilibrium** when all the forces and torques balance out. However, even if the object is in equilibrium it may still suffer deformation as a result of these forces. The amount of deformation is quantified by **strain** and the combination of forces causing this strain is called **stress**. We will talk a bit about these concepts in Lecture 13 on elasticity. As you can guess, the details can get pretty hairy so we will only cover the high points.

For now we will ignore all of these issues. If we focus our attention on physical systems with negligible configuration (i.e., a particle) then there is no deformation to consider. In fact, the system has no extension, so the idea of torque doesn't even apply. All we are left with is the simpler notion of force and its effects.

Obviously an object subject to a single force cannot be in equilibrium. We need two or more forces and they need to balance. Two forces must be equal and opposite in order to create equilibrium. In vector notation we may say:

$$\vec{F}_1 + \vec{F}_2 = 0$$

Because of the way we have defined vectors in Lecture 2, this same formula works for an arbitrary number of forces:

$$\sum \vec{F}_i = 0$$

Here the summation is implied to run over all the values of i. If there are four forces then the summation is a shorthand for

$$\vec{F}_1 + \vec{F}_2 + \vec{F}_3 + \vec{F}_4 = 0$$

If necessary we may put a subscript on the summation symbol for clarity:

$$\sum_i \vec{F}_i = 0$$

or even more explicitly:

$$\sum_{i=1}^{4} \vec{F}_i = 0$$

But I'll usually be sloppy about this notation unless it is likely to cause confusion.

We walked through a typical force equilibrium problem in Lecture 2. It might be worth reviewing that example now (see page 11).

There are also different kinds of equilibrium: stable versus unstable. When forces vary in time, the equilibrium they create may be destroyed. Typically the forces in a system depend upon its configuration (usually the distance between its parts). So, any motion in the system can change these forces. If these forces are such that any displacement tends to cause the system to return to equilibrium, this is **stable equilibrium**. A simple example is a marble in a bowl. If we displace the marble away from the center of the bowl it will be pushed back to center. Turn the bowl upside-down and you have **unstable equilibrium**. A pencil standing on it's point is another example. You may have seen a person holding a broom upside-down with the palm of their hand. This is an example of **dynamic equilibrium**. In this case the forces are not dependent on the configuration but on a feedback loop (i.e., the person doing the balancing). Dynamic equilibrium is often used to combat unstable equilibrium, but if one is not careful it can lead to literally explosive results.

For now we only consider forces that are constant. So if they are in equilibrium they will remain that way.

There is one other point to make regarding equilibrium. No net force on a object does not imply that the object is stationary. What it implies is that the acceleration is zero—the object may be moving with constant velocity. We saw an example of this with air drag and terminal velocity in Lecture 3. Although it is not common to make this distinction, it may be worth calling equilibrium on a moving object **kinetic equilibrium** as opposed to **static equilibrium** for an object at rest.

In our inventory of mechanical forces to consider, the simplest is **weight**.[1] In the English system of units the pound is a primary unit. However, the influence of gravity can vary slightly depending on where you are on the surface of the planet. This means that the weight will vary as well. In the metric system, mass is primary and weight is secondary. Mass is a difficult quantity to define precisely without using Newton's laws (to be discussed in the next lecture). However, the weight of an object is simply related to its mass via

$$W = mg$$

If the gravitational acceleration increases slightly so does the weight. The SI unit for weight is called the **newton** and the metric unit for mass is the **kilogram**. Under standard earth gravity, a one kilogram mass will weigh about 9.8 newtons which is equivalent to about 2.2 pounds.

[1] Our book discusses weight in the context of the Newton's law of gravity (section 4.7). I prefer to wait and talk about the law of gravity in Lecture 6.

Tension will come up again when we discuss elasticity in Lecture 13, but for now we only consider it in the context of an **ideal string**. A string is ideal if it has negligible weight and does not stretch at all. A string like this essentially transmits the force from one end to the other. Combine this with an **ideal pulley** (frictionless, negligible mass) and you can create a block-and-tackle. Take for example Figure 4.1. Notice how the larger weight is supported by two strings. But the two strings are really the same string, so the force that supports the weight is actually the tension multiplied by two. The tension is in balance with the smaller weight. This block-and-tackle system essentially multiplies force by two. This is called the **mechanical advantage** of the system. Actually, this is the ideal mechanical advantage. The true mechanical advantage will take into account the friction and the masses of the pulleys and strings. In fact, the ratio of actual to ideal mechanical advantage is called the **efficiency** of this simple machine.

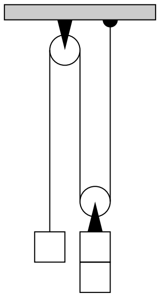


Figure 4.1: Simple block and tackle with mechanical advantage of two

The third mechanical force we will consider is a **support force**. Put a 10 kilogram mass on the table. The table supports this mass by holding it up against its weight. This force of support is ultimately from elastic forces: the weight actually deforms the surface of the table slightly. The table resists this deformation with a support force. Understanding this is important because there is no "formula" that governs the support force—it is a reaction to the other forces in the problem. The magnitude of the support is simply that which is required to balance the forces against it.

These forces are sometimes called **constraint forces** because the surface that provides the support constrains the motion of the object. The motion can only occur parallel to the surface because the force resists any motion into the surface. I prefer to call them support forces because that seems to me to describe them best. However, it is much more common to call these **normal forces**. The reason for this is that in mathematical jargon the word "normal" means perpendicular and these forces always operate perpendicular to the surface doing the supporting.

A more involved example of the use of support forces in a problem is a weight supported by an inclined plane. Review Figure 4.2. Because the support is perpendicular to the slope of the plane, only a component is available to counterbalance the weight. The remaining component wants to push the block down the slope. However, the tension in the string tied to the smaller mass holds it back. This tension is coming from the weight of the smaller mass pulling down across the pulley. If the masses are in the right combination, these forces will balance in equilibrium.

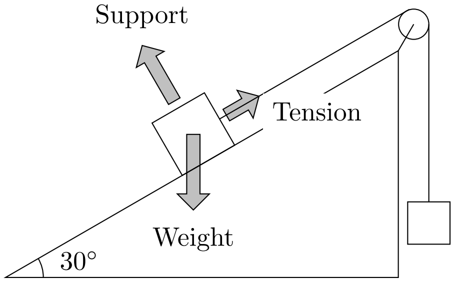


Figure 4.2: Inclined plane with pulley

I have redrawn these three forces in Figure 4.3. When we collect all the forces acting on one part of the system, it is called a **free-body diagram**. It is almost always best to align your coordinate system with any forces of constraint. Usually the constraints are the most difficult forces to calculate in a problem. By properly orienting the coordinates we can usually begin the calculation in the dimension perpendicular to the constraint to get more information before tackling the constraint itself. In fact, it is sometimes possible to solve a problem without even solving for the constraint.

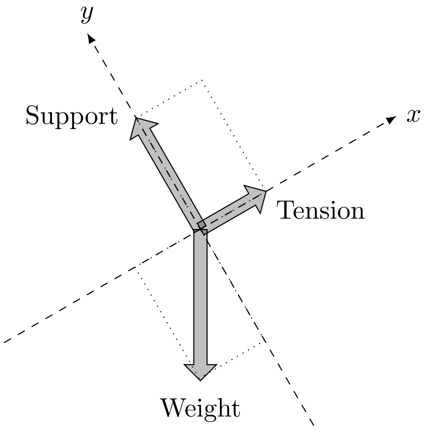


Figure 4.3: Free-body diagram from Figure 4.2 (magnified 2x)

In this case it is pretty simple. Notice how the components of the weight correspond exactly with the tension and support forces. This shows that the masses are in equilibrium.

If we call the larger mass #2, the component of its weight that points down the plane to the left is given by

$$W_2 \sin 30^\circ = 0.500 W_2$$

This must equal the tension in the rope and that tension equals the weight of the smaller mass (which we will call #1). We have $0.500W_2 = W_1$. The ideal mechanical advantage of this inclined plane is two. This can be seen by

$$MA = W_2/W_1 = 2$$

In general, the ideal mechanical advantage of the inclined place is given by $1/\sin\theta$.

The fourth mechanical force to consider is friction which occurs whenever two surfaces are in contact. The amount of friction depends on two factors: (1) the amount of support force that is pressing them together, and (2) the nature of the two surfaces in contact. The effect of the second factor is captured in the **coefficient of friction** and is given the symbol μ. Its value is between 0 and 1. The formula for friction is[2]

$$F = \mu N$$

Now, there are two varieties of friction: static and kinetic. If the block is sliding, **kinetic friction** applies and the force is given by $F_k = \mu_k N$ where μ_k is the coefficient of kinetic friction. Remember that the forces on a moving object may still be in equilibrium (previously we called this kinetic equilibrium). You may encounter a homework problem or two where an object is sliding at constant speed. In that case, you know two things: (1) the friction is kinetic and (2) the forces all balance.

If the block is not sliding, then clearly the forces are in static equilibrium. In this case the **static friction** is whatever it needs to be to maintain balance—just like the support forces we studied earlier. However, the static friction has a maximum value. If the force required to maintain balance exceeds this maximum value, the block won't be able to resist moving. Thus, if the block is stationary we know the static friction is less than this maximum value. This upper bound is given by the formula $F_s \leq \mu_s N$, where μ_s is the coefficient of static friction—a number greater than its kinetic cousin.

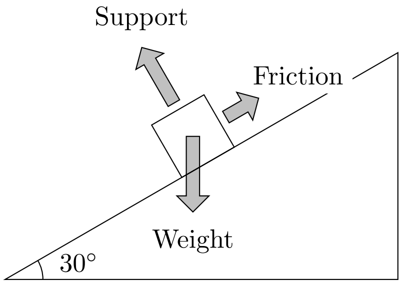


Figure 4.4: Inclined plane with friction

As an example, consider Figure 4.4, which is essentially the same as Figure 4.2 with the tension replaced by friction. We will assume that this object is in equilibrium, so the free body diagram is essentially the same as Figure 4.3.

Therefore the magnitude of the normal support force is equal to

$$W\cos 30^\circ = 0.866W$$

and the magnitude of the friction force is equal to

$$W\sin 30^\circ = 0.500W$$

Since the static friction force is related to the support force according to $F_s \leq \mu_s N$, we have a constraint placed upon the coefficient μ_s:

$$\mu_s \geq 0.577$$

In other words, if the coefficient of static friction is less than 0.577 (in general, $1/\tan\theta$), the block will slide because there won't be enough static friction to hold it in place.

This is the last constant mechanical force for us to consider. However, we will see more forces in the upcoming lectures. For example, in Lecture 13 we will discover the formula for elastic forces. Truly, elastic forces are the cause of both tension and support, but in that lecture we will focus on the role of deformation in elasticity which we are neglecting here.

All of these forces that I've classified as "mechanical" (including elastic forces) require the objects to be in contact. In contrast, gravity and the electromagnetic forces are **long-range** forces which act at a distance. We will talk about the electric and magnetic forces in Lectures 23 and 24, respectively. The law for the electric force is quite similar to Newton's law of gravity which we will study in

[2] This is our first example of an **constitutive equation**. I mentioned in the first lecture about how one of the roles of the deductive approach is to explain from first principles the values obtained for equalities such as these. This one, however, is nearly impossible and is dependent on a variety of factors. As such, these numbers for various combinations of surfaces can only be derived on average from the lab.

Lecture 6. The magnetic force is unique in that it causes a deflection that is perpendicular to the motion of the particle. Using results from Lecture 6 we will see that this produces a characteristic spiral motion. This perpendicular deflection is also a characteristic of the Coriolis force we will mention in Lecture 9.[3]

In Lectures 29 and 30 we will also encounter the existence of two nuclear forces. This is really a misnomer though because at that level we cannot avoid using quantum mechanics and the idea of force must be replaced with the more general notion of "interaction". We will lay the ground work for moving beyond force in Lectures 7 and 8 when we talk about energy and momentum.

Next week we will introduce Newton's three laws of motion. The first law will cause us to consider reference frames in motion. We will find that if we are not careful choosing these frames we may introduce "fictitious forces" into the laws of motion. We will also see where Einstein's relativity touches the laws of motion. Newton's second law establishes the connection between force and acceleration alluded to in Lecture 3. Finally, the third law will introduce us to systems and the role of internal and external forces.

[3] There is a peculiar fascination with the distinction between contact forces and long-range forces in the history of physics. The long-range nature of gravity really bugged Newton: he called it "occultish". Prior to Newton, it was felt that any force had to be a contact force: for example, Descartes filled the solar system with a fluid that kept the planets in motion. As a consequence of the study of electromagnetism in the 19th century, a kind of middle ground was established: the force field. We say that space is filled with gravitational and electromagnetic fields which operate as media for the transmission of these forces. But the field is not mechanical—they obey laws of their own independent of mechanics. However, the modern development of quantum field theory has written a new chapter in the debate between long-range and contact forces which comes nearly full circle.

Lecture 5

Newton's Laws of Motion

Read sections 4.1–4.5 and 4.11, sneak a peak at sections 28.1–28.2

Inertia is the property of a system to maintain its state of motion. When an object is moving it has a tendency to keep moving. This is why you never want to stand in front of a moving train. This principle of inertia may seem obvious but it has not always been that way. The critical modern insight is to set aside the realities of friction and weight. Without this conceptual distinction, Aristotle taught that every object has its place. In other words, objects have an intrinsic tendency based on their composition to move until they come to rest in their place. Things made of earth fall down, things made of air float up.

Galileo was the first person to apply the idea of inertia to objects in motion in a quantitative way. He was able to see the central point that once an object is set in motion the only reason it stops is friction and that friction is external to the object. Whether it is dragging against air resistance or dragging across a surface, if this friction is eliminated there is nothing to stop the motion. In situations where the effect of friction is minimized (an icy road, a rolling ball) this inertia is more obvious.

This principle of inertia is quantified in the notion of **mass**. If an object is in motion, the larger mass is harder to stop. This works in reverse too: the more massive an object, the more difficult it is to push it into motion. If we decide on a particular object as our unit of mass we can measure the force required to get it to move at a particular rate. If twice the force is required to get a second object to move at the same rate, we say it has twice the mass. So mass measures the amount of resistance an object will have to changing its state of motion.

Newton's first law of motion is Galileo's principle of inertia. Newton's wording is as follows:

> Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon.

When first exposed to this idea it is easy to associate it with something like pushing a box across the floor: when you start out, with just a small amount of force, the box does not move at all. Only after a certain critical level of pushing has been reached will the box move. This is not inertia. This is friction. It is better to think of the force required to stop an object in motion—this is tied to the idea of inertia more directly than considering the force required to start the motion.

Galileo also recognized something we now call the **principle of relativity**.[1]

[1] Not the theory of relativity—that was Einstein. See Lecture 26.

Imagine you are watching boats sail across a calm lake. There is a man hanging off the mast of a ship making repairs. The wrench he is holding slips and drops 30 feet straight down striking the base of the mast (we are going to ignore air drag). Since the wrench has a bright orange handle you see it clearly tumble down as the ship drifts across your field of view. But from your viewpoint, the path of the wrench is not straight but a parabola. Of course you agree that the wrench strikes the base of the mast, but to you the ship just happens to catch it on its way down.

Well, "just happens" is an overstatement because the reason the wrench is not falling straight down is that before it was dropped, it was moving with the drift velocity of the ship. The wrench continues to drift with this velocity while it undergoes acceleration downward due to gravity.

Here's the point: is the ship moving? Yes, you say, of course it is moving: it is drifting across the lake. What does the repairman say? Yes, he says, I can look out at the shore and see the ship is moving. Fine: everyone agrees. Later, below deck, he is eating lunch and drops his fork. The fork falls straight down like the wrench—is the ship moving? Yes, but without a reference frame tied to shore, the repairman cannot tell. All motion is always measured relative to some reference frame. And this is the principle of relativity: within a reference frame moving with constant velocity, it is not possible to determine this drift velocity based on the relative observations made within that frame. The laws of physics are independent of the velocity of the frame.

This abstract symmetry is broken by the fact that we live on earth. There is a distinct up-and-down direction due to gravity and a distinct state of rest due to friction and air drag. But take away gravity and friction and you have this kind of free-fall world where the principle of relativity rules.[2]

There is a positive way to state the principle of relativity, too. Since the laws of physics work in any frame (moving with constant velocity), you are free to choose whichever you like as your rest frame. In fact, in order to do any kind of calculating, you must start by choosing this frame. The point is that the choice is arbitrary—there is no "natural" choice.

But not completely arbitrary. It may seem like the airplane is at rest when you are at cruising speed, but hit some turbulence and you're not at rest any more. The key is that the frame must move with constant velocity: no acceleration. A reference frame in which Newton's first law holds is called an **inertial frame**. Non-inertial frames like a car making a turn will have pushes and pulls which are due to the non-constant velocity (either speed or direction) of the frame. We use Newton's first law to identify an inertial frame. The principle of relativity implies that any frame moving relative to another inertial frame with constant velocity is also inertial.

Newton's second law of motion presumes we are working in an inertial frame defined by the first law. According to Newton we have:

> The alteration of motion is ever proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.

This is commonly written as

$$\vec{F} = m\vec{a} \tag{5.1}$$

This formulation is due to Euler.[3] Notice how neither acceleration nor mass are mentioned in Newton's formulation. He uses the phrase "alteration of motion" which may seem vague, but earlier he defined this "quantity of motion" as:

> The quantity of motion is the measure of the same, arising from the velocity and quantity of matter conjointly.

[2] This is one of the reasons why a pool table is such a good place to learn physics. The horizontal nature of the table eliminates the effect of gravity and the balls roll so the effects of friction are minimized.

[3] Euler's contributions to both mathematics and physics are extensive. In a sense Euler completed Newton's laws by extending them to rigid objects—cf. Lectures 9 and 10.

Today we use the term mass for "quantity of matter" and the "quantity of motion" we call momentum (see Lecture 8). In other words, Newton's second law as originally formulated states that force changes momentum. Now if the mass of the object is constant, the part of the momentum that changes is the velocity, so this is equivalent to Euler's formulation. But if the mass is not constant (a rocket ship is a good example) the more general law based on momentum is needed.

Unless otherwise stated, assume the mass is constant and use equation (5.1).

If an object is under the influence of several forces, we need to calculate the vector sum of these forces to determine the direction of acceleration. Clearly the object can only move in one direction, so we need one net force vector to put on the left side of the second law. This is often the most difficult part in solving these problems: identifying the forces and adding them up to some net force. Of course, the forces may all balance. In this case the vector sum is zero. Accordingly equation (5.1) implies $\vec{a} = 0$, so the object in equilibrium will either move with constant velocity or remain at rest. You can see how the problems from the previous lecture are now a special case of the more general second law.

Another way to express Newton's second law is that any unbalanced force will produce acceleration. This emphasizes the initial step of calculating the net force on an object. A classic example problem is the **Atwood machine** which is simply two weights hanging over a pulley (see Figure 5.1). Just by looking at the diagram you can see what will happen: the larger weight will drop but slower than normal because it will have to pull the smaller weight up. We now need to calculate the result.

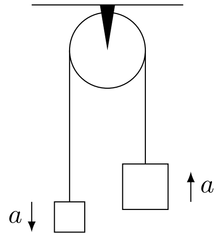


Figure 5.1: Atwood machine

One important thing to realize here is that this is a system with two parts (three if you count the string). Each weight has a different net force acting on it and each weight corresponds to an application of equation (5.1). Let's call the mass of the smaller weight m and the mass of the larger weight M. Therefore the weight of each is mg and Mg respectively.

The force that opposes the weight of each is the tension in the string, which we will label T and it is equal on both sides. But recognize that the tension is undetermined—it is equal to neither the smaller nor larger weight. This is an easy mistake to make. I think this is because we draw the diagram and forget that it represents a snapshot in time—the objects are actually moving and the forces are out of balance. I try to draw a little arrow with an a to remind myself the objects are accelerating.

So, we are ready to apply Newton's second law. This problem is simple so we will jump the result. We'll walk a bit more slowly in the next problem to show all the steps. From the smallest weight we have:

$$T - mg = ma$$

and from the larger weight we have:

$$T - Mg = -Ma$$

Notice the signs. You must be deliberate with the signs because these quantities are vectors. In both cases the tension from the string pulls up, so they get positive signs. Similarly the weights pull down, so they get negative signs. The smaller block will accelerate up so it is positive, and the larger block will accelerate down so it is negative. The magnitude of the acceleration of the two blocks are the same because they are tied together.

From a mathematical standpoint we have two equations with two unknowns (T and a). This means we can solve for both. In order to solve for the acceleration we will use the first equation to eliminate the T from the second. This yields

$$(ma + mg) - Mg = -Ma$$

Solving for a gives

$$a = \frac{M - m}{M + m}\, g$$

It is sometimes helpful to look at some extreme cases as a double check. If $m = M$, then we have $a = 0$ which makes sense—the weights balance. If $m = 0$ then $a = g$ which also makes sense because if there is no smaller weight the larger will simply fall under the influence of gravity. Intermediate cases are in between, so this answer checks out.

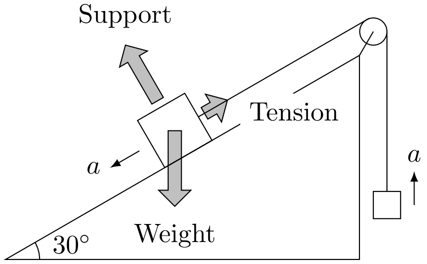


Figure 5.2: Inclined slide with pulley

Now for a more complicated example. Examine the situation in Figure 5.2. This is similar to Figure 4.2, but the hanging weight is smaller (and the tension is less) so the larger mass will slide to the left. Suppose the larger mass is 10 kilograms and the smaller mass is 3 kilograms. How long will it take the block to slide down the incline from rest if the plane is one meter long? What will be its final velocity?

The first step in any of these problems is to draw a good free-body diagram for each piece of the system. Fortunately for me we have already done this in the Lecture 4 (see page 23) except that the tension is a bit smaller now. In this problem the tension is not equal to the smaller weight because the forces are not in balance. In fact we know that the tension is larger than the smaller weight because it will need to lift the smaller weight up as the larger mass slides down the incline. The situation on this smaller mass is similar to the previous example, so we have

$$T - mg = ma$$

just as before. This gives us a formula for tension, $T = m(a + g)$, which we can use in the analysis of the motion of the larger mass.

Looking back at the free body diagram in Figure 4.2, we can see that the tension opposes the x-component of the weight. We know that these will not balance, but we can still use Newton's second law. In this case, the weight is 98 newtons. The x-component of the weight in this inclined coordinate system is

$$W_x = (98)(\cos 240°) = -49$$

The 240° is measured from the positive x-direction. The sign of this component is negative because it points to the left. Notice how the positive x-direction for our coordinate system is pointing up the incline. This also means we should expect the acceleration of the larger block to be negative.

So, from the x-component of the larger mass we have

$$-49 + (3)(a + 9.8) = (10)(-a)$$

which we can solve for a. Thus,

$$a = 1.5077$$

We have not yet answered the question, however. We were asked about time and final velocity. We were given the initial velocity ("at rest") and the displacement, so we should consider using equation (3.4) to get the time. Thus,

$$(-1) = (0)(t) + \tfrac{1}{2}(-1.5077)(t)^2$$

The negative signs are there because the motion is to the left. The final answer is 1.15 seconds. In order to determine the final velocity, we use equation (3.6). Solving yields 1.74 meters per second.

The various combination of possible problems are endless. Add friction into the mix and you have a really hard problem. But in each case the procedure is the same. Focus first on getting the acceleration of the objects by applying Newton's second law. Thus,

1. Identify all the forces in play.
2. Draw a free-body diagram for each part of the system (remember to align your coordinates along any natural constraint).
3. Determine the components of the net force in each diagram.

Each part and each dimension will yield an equation via Newton's second law. Sometimes you won't need all of these equations but they are all valid. Once you have the formulas, calculate the acceleration of the objects.[4] From there use the techniques from Lecture 3 to answer the specific question.

So far we have used Newton's law to calculate acceleration given a certain set of forces. The process can work the other way too. We can investigate the character of a force from the acceleration it causes on the objects it influences. For example, we can do experiments with static electricity do determine the way that the electric force is dependent upon distance.

One interesting example of using the second law this way is in dealing with **inertial forces**. I have said before how the second law is only valid in an inertial frame. So a particle free from the influence of outside forces will move in a straight line at constant speed. If we are in a non-inertial frame, the frame will move underneath the force-free particle. From the point of view of someone tied to the frame, the particle will experience an acceleration. Being a good student of Newton, this person would be led to postulate the existence of some force that causes this acceleration. However, this force is not due to outside forces but is an artifact of the non-inertial frame. This is why these kind of "forces" are called fictitious forces.

The most common example of such is the **centrifugal force**. If you stand on a merry-go-round you will feel a force that wants to fling you off—out from the center of rotation. Your inertia is propelling you in a straight line which, relative to the rotating merry-go-round, is out. You must overcome your inertia by pushing in to stay on the rotating frame. You interpret this as counter-balancing some centrifugal force. But this centrifugal force has no source: no object is pulling you out. It is merely a manifestation of the principle of inertia.

These inertial forces are always proportional to the mass of the object. This is because they are enforcing the principle of inertia. The greater the mass of an object, the larger its inertia. These forces always take the form ma from equation (5.1), so that the resulting acceleration is independent of the object. This is because the acceleration is due to the nature of frame rather than the object. Another way to recognize an inertial force is that it will disappear if the right frame of reference is chosen. These frames are precisely the inertial frames we talked about earlier.

Notice that gravity has these same qualities. We infer the force of gravity from the acceleration we see in falling objects. This acceleration is independent of the object. The force of weight is given by the formula mg. The fact that the mass which appears in the weight formula is the same as the inertia that appears in Newton's second law is called the **equivalence principle**. When Einstein went looking for a place to rebuild a new theory of gravity consistent with relativity, he started with this principle. He imagined a physicist in a falling elevator. In this frame, gravity would appear to be turned off. Therefore, a reference frame accelerating in free-fall is actually an inertial frame of reference. This thought experiment only works for a small space (for instance, you can't find a frame that will "turn off" gravity for the whole planet), but that was enough of a starting point to build the general theory of relativity.

Newton's third law of motion is often a bit confusing because the context is different than the first two. In the first two laws we deal with a single particle. Even when we used the second law to analyze a system, the second law only applies

[4] If the system has multiple parts usually the accelerations are the same, but sometimes not—the parts of a block and tackle systems will not have the same acceleration, for example.

for one part at a time. The third law deals with pairs of particles. Newton sez…

> To every action there is always opposed an equal reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

Until now we have discussed forces that come from outside. Now we talk about forces internal to the system of particles—forces of interaction. The third law is telling us that every interaction in nature involves an equal and opposite pair of forces. It is a statement about the nature of force itself.

In practice, the third law is not very helpful. It's one of those things that make the whole conceptual structure work, but the hard work is done by the second law. It does imply something that is very important however. Since every force changes momentum and every interaction involves equal and opposite forces, every interaction involves equal and opposite changes in momentum. In other words, we can conceive of the interaction as consisting in the exchange of momentum. This viewpoint forms a perfect compliment to the idea of energy and we will explore this perspective more in Lectures 7 and 8.[5]

But for now we will stick with Newton's laws for a bit. Next week we will explore our first non-constant force: the force required to maintain circular motion. We will talk a bit about circles and the velocity and acceleration involved in circular motion. This will lead into the idea of centripetal force. We will discover that the force of gravity which holds the planets in place is also centripetal. Newton's law of gravity will be introduced and we do some prep work for the Lecture 11 on celestial mechanics.

[5] It is also a key element of quantum field theory—see Lecture 27.

Lecture 6

Circular Motion and Gravity

Read sections 5.1–5.5, review section 4.7, and sneak a peak at sections 10.1–10.2

The ancient Greeks had a love affair with circles and spheres. These are the only shapes that can change without changing. Looking out at the night sky, they imagined astronomically large transparent spheres spinning around the earth. In each was lodged a planet which rotated with them like a embedded jewel.

Though we no longer think that way about the night skies, there is still something magical about the circle. We can talk about the paradoxical ratio π, the complications of abstract geometry, or complex numbers—behind every cycle in nature is a circle waiting to be found. We will exploit this fact in Lecture 12 when we talk about simple harmonic motion. But for now we will talk about some basics.

The most important point of the circle is not on the circle: its center. The circumference is related to the radius of the circle by $C = 2\pi r$. So the average speed of an object rotating in a circle is given by $v = 2\pi r/T$ where T represents the amount of time required to complete one cycle which we call its **period**. We will call the motion **uniform circular motion** if the speed has this average value the whole time.

We will normally talk about uniform circular motion in the context of a full circle, but it does not have to be that way. A car making a right turn can move with uniform circular motion for a quarter of a circle. The results from this lecture will be applicable to that motion as well. Whenever a path is curved that curvature can be characterized by a circle with the same curvature. Since the circle is characterized by its radius, we can quantify this curvature by its **radius of curvature**. The smaller the radius of curvature, the tighter the curvature of the path.[1]

We know from the principle of inertia that circular motion cannot be maintained without a force. It would be helpful to know what acceleration is required to maintain a particular state of uniform circular motion. The formula

$$a = v^2/r \tag{6.1}$$

can be derived by comparing the velocity vectors at two different points in the motion. See Figure 6.1. By definition, the acceleration is the difference between these two velocity vectors divided by the time it takes to get between them. These three vectors form a triangle. In the meantime, the radius vector that sweeps from one point to the other also forms a triangle. The other side of this triangle is almost the same as the arc-length of the circle traveled.[2] These two triangles are similar, so the ratio of the sides are equal. In particular

$$\frac{at}{v} = \frac{vt}{r}$$

[1] In fact, the curvature is related to the reciprocal of the second derivative of the path. One way to see this is that the first derivative is related to the slope or direction of the path. The second derivative will quantify how much this slope or direction changes.

[2] This approximation becomes better the smaller the duration involved.

From this formula equation (6.1) follows directly. These considerations also show the direction of the required acceleration: toward the center of the circle.

So any force that is going to produce this motion will need to be equal to mv^2/r. You will often see the equation

$$F_{\text{cent}} = \frac{mv^2}{r} \tag{6.2}$$

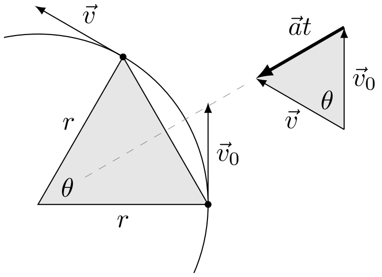


Figure 6.1: Acceleration required for uniform circular motion

There is nothing wrong with this equation, *per se*. It just gives the wrong impression like we have discovered a new fundamental force. No, what this formula represents is the net force required to maintain uniform circular motion. You see, we are working backwards relative to the work we did in the previous lecture. Previously we had the forces and inquired concerning the motion. But now we have a preconceived motion and want to know the constraint on the net force producing this motion. For uniform circular motion this net force is called **centripetal force**.

For example, suppose we have a twisting slide and we want to know at what angle to bank the curve. Let's also suppose the typical velocity of the person on the slide is 10 meters per second and we want the radius of curvature to be 5 meters. In this problem there are two forces: the weight of the person and the normal support force from the slide. See Figure 6.2. Where is the centripetal force? It is the net force from these two that causes the circular motion.

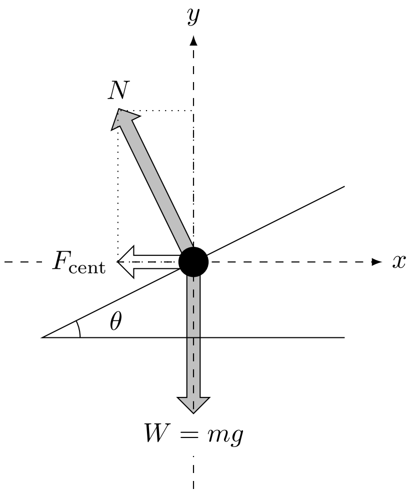


Figure 6.2: Slide banked to produce uniform circular motion

In this problem, the centripetal force is coming from the x-component of the support force. The y-component balances the weight. How do I know they balance? Because the net force is horizontal. I have aligned the coordinates with the motion of the system.[3]

The angle the support force makes to vertical is the the same θ that the incline makes to horizontal. We know the y-component of the support is mg, and the two components are related to the tangent via

$$\tan\theta = N_y/N_x$$

So,

$$N_x = \frac{mg}{\tan\theta}$$

The whole point was to set this horizontal force equal to mv^2/r. Therefore

$$\frac{mv^2}{r} = \frac{mg}{\tan\theta}$$

Notice how the mass cancels from both sides, so the angle of incline does not depend on the mass. We know all the rest of these variables, so we can solve for θ. We get

$$\theta = \tan^{-1}(gr/v^2) = 26.1°$$

As another example, consider an ion traveling in a magnetic field. We will find in Lecture 24 that there is a magnetic force proportional and perpendicular to its velocity. Because the force is perpendicular it deflects the path without changing its speed. The path of particle is circular. The magnetic force is given by $F = qvB$ where q is the charge of the ion, v is its speed, and B is the strength of the magnetic field. The radius of the circular path can be determined by setting this equal to mv^2/r and solving for r. One gets

$$r = \frac{mv}{qB}$$

If the particle comes in at an angle oblique to the magnetic field, the component of the velocity parallel will be unaffected so the particle actually spirals up or down the field.

[3] This may look like an exception to the rule of aligning the coordinates with the constraint. Perhaps it is, but the point is you want to align them in order to make the math simple. In this case it is simpler to align it with the centripetal force. You can set the coordinates up however you want. Sometimes you can't help it but usually you can set things up to avoid a bunch of spaghetti math.

This happens in the earth's magnetic field. The solar wind is composed of ions of hydrogen. When they enter the earth's magnetic field they begin a spiraling motion and are funneled along the magnetic field lines. As they get closer to the poles, the magnetic field becomes more dense and the radius of the spiral decreases. The collision of these particles with the atmosphere causes the Aurora Borealis, more commonly known as the northern lights.

We know that every particle in uniform circular motion does so because it is under the influence of a centripetal force of some sort. This also applies to the solar system. Now, even to the ancient Greeks it was known that the planets do not orbit around the sun in a perfect circle.[4] But once the transparent spheres were definitively shattered by Galileo, the nature of the centripetal force holding the planets in place came into question. By Newton's time it was a common assumption that this force obeyed an inverse-square law. That is, the force of gravity between two objects is inversely proportional to the distance between them. What Newton did was to prove it.

The application of Newton's laws of motion to the motion of the the planets dominates the *Principia*. Along with his three laws of motion, he lays out the universal **law of gravity** as

$$F = \frac{GMm}{r^2} \tag{6.3}$$

We will limit ourselves to the special case when $M \gg m$ though Newton himself did not. This means that the larger mass M is basically unaffected by the motion of the smaller mass m. The value of G is a universal constant characterizing the force of gravity. In SI units it has a value of 6.673×10^{-11}.

For Newton the force of gravity was the archetypal centripetal force. However, for him the term meant something different than it does for us. We reserve the notion of centripetal force for that associated with circular motion. For Newton a centripetal force was any force whose direction was always toward a particular point in space. Today we would use the term a **central force** to describe such a thing. So we say that gravity is a central force.

For any central force it makes sense to describe motion in terms of **polar coordinates**. In polar coordinates we describe the location of an object by its distance from a particular point in space and its angle from a particular line from that point. This is an example of a non-Cartesian coordinate system that can be helpful in the analysis of orbital mechanics. Compare the systems in Figure 6.3. We will mention this again in the Lecture 11.

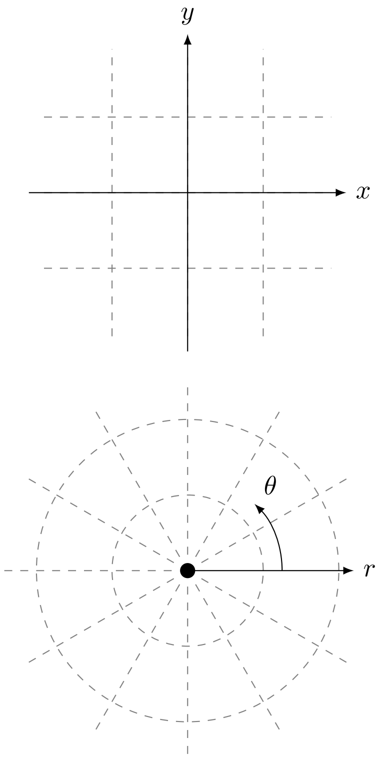


Figure 6.3: Cartesian versus polar coordinates

A generation before Newton, Kepler had painstakingly examined the best known astronomical data of his day (mostly from the study of the motion of Mars). He was able to distill their pattern into what is now known as **Kepler's three laws** of planetary motion. They are:

1. The orbit of the planets are elliptical with the sun located at one of the foci.
2. The rate of area swept out by the line from the sun to the planet is constant.
3. The square of the orbital period is proportional to the cube of the radius of the orbit.

It is easy to show that Kepler's third law is a consequence of the law of gravity by simply setting the inverse-square law equal to mv^2/r. Thus,

$$\frac{GMm}{r^2} = \frac{mv^2}{r} \implies GM = rv^2$$

But the orbital period is related to the velocity by $v = 2\pi r/T$, so we have

$$GMT^2 = 4\pi^2 r^3 \tag{6.4}$$

[4] Mercury and Mars have the least circular orbits of the eight planets in the solar system.

This is Kepler's third law.[5] Newton was also able to show that the other two laws followed from the law of gravity, but those proofs are a bit more complicated.[6]

Another example of the use of these formulas is in calculating **geostationary orbit**. This the distance in which a satellite will have an orbital period of exactly one day. This means that the satellite will simply remain overhead and is useful for communication satellites. Since there are 86,400 seconds in a day, we can calculate this orbit distance using Kepler's third law. For the earth we have $GM = 3.99 \times 10^{14}$, so

$$r = 4.22 \times 10^{7}$$

or 42,200 kilometers from the center of the earth. This corresponds to an altitude of 35,800 kilometers above sea level. GPS satellites orbit at an altitude of 20,200 kilometers so they orbit about twice a day.

Historically one of the great paradigm shifts in science occurred when Newton tied the concept of universal gravity to the common notion of weight. We can apply equation (6.3) to the surface of the earth. We know that the radius of the earth is 6380 kilometers. Therefore the force of gravity on the surface of the earth is

$$\frac{GMm}{r^2} = mg \implies g = \frac{GM}{r^2}$$

Plug in the numbers and you get our familiar 9.8 meters per second squared. This allows us to calculate the acceleration due to gravity on other planets. On the moon it's 1.6 meters per second squared. In this way Newton was able to unite motion in the heavens with motion on the earth.

Next week we will introduce the idea of energy. The concept is motivated by a study of the force-multiplying power of machines. We will see that, if we ignore friction, energy is a conserved quantity. The advantages of having a conserved quantity will allow us a much easier way to solve certain mechanical problems. We will find that energy comes in three basic forms: kinetic, potential, and heat. Finally, we will introduce the idea of least action.

[5] In 1934 the analysis of the rate at galaxies spin was determined to violate Kepler's third law. The speed of the stars on the outer edges is much too fast. This implies that there is a lot more mass in the galaxies holding them together than can be inferred from the stars we see. This and other astronomical observations lead many to postulate the existence of **dark matter** in the universe (another way of dealing with the problem is to assert that the inverse-square law of gravity does not hold for these extreme distances). It's a bit embarrassing to admit that we have no clue about the composition of approximately 80% of the known universe. See here for more info.

[6] In Lecture 11 we will see how the second law follows from the central nature of the force of gravity

Lecture 7

Energy and Action

Read sections 6.1–6.9

Back in Lecture 4 we touched on the idea of mechanical advantage. The purpose of any mechanical machine is to multiply the input force in order to create a much larger force for useful work. It is possible to break the analysis of a machine into components each connected together. These components are called **simple machines** and are traditionally classified as

- Lever
- Wheel and axle
- Pulley
- Inclined plane
- Wedge
- Screw

This list could be reduced to two: the lever and the inclined plane. The first three all operate based on a twisting motion around a pivot, while the second three operate based on splitting the support force that counter-balances a perpendicular force. We've already talked about the inclined plane on page 23. We will talk about the lever again in Lecture 10 with the idea of torque, but we will soon see in this lecture a different way to analyze the problem.

The multiplication of force from these machines is not without a cost. The cost is displacement. Refer back to the block-and-tackle system in Figure 4.1 for a moment. This system is in equilibrium, so the smaller block is supporting twice its weight. Now imagine you were to pull down the smaller block by 10 millimeters. Because of the way the pulleys and strings are set up the larger block will rise only 5 millimeters. So, the mechanical advantage of two corresponds to a two-fold reduction in the distance we can raise the weight. This implies that there is some sort of conservation happening here. The product of the weight and displacement is the same on the input and output (ignoring friction). This product we define as **work**.[1] The SI unit is called the **joule** and is equivalent to the newton multiplied by the meter.

Now consider the lever in Figure 7.1. The three blocks on the left are balanced by the one block on the right. But notice that the block on the right is three times the distance from the fulcrum as the three blocks on the left. Because of this, when the three blocks move down one millimeter the lone block will move up three. So a force that is on the right side must move the lever three times as far in order for its force to be multiplied three-fold on the left. Again the work is equal on both sides. This approach of analyzing the forces in a system by determining

[1] Be careful not to confuse this technical definition with its looser English equivalent. Work is not synonymous with effort. In particular, if there is no movement there is no work done.

how a small displacement propagates through the system is called the principle of **virtual work**.

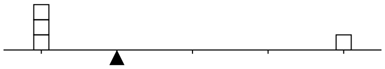


Figure 7.1: A simple lever

Both force and displacement are vectors, so we need to say what we mean in the general case where they point in different directions. We only want to have the component of the force that contributes to the displacement considered. If θ represents the angle between the two vectors then we want to use the formula[2]

$$W = Fd\cos\theta \tag{7.1}$$

Notice that work can be negative if the force opposes the displacement.

You may remember the dot product I mentioned in Lecture 2. We can use that to express this definition as $\vec{F} \cdot \vec{d}$. Note in particular that when the force and displacement are perpendicular, the work done is zero because that force does not contribute to the motion.

So for any simple machine the work done by the input force is equal to the work done by the output force. Chain them together and the same is true for a more complex compound machine. But there are always friction effects. The ratio of the output work to the input work is equivalent to what we have previously defined as the **efficiency** of the machine.

The ability to do work is a valuable quality in any machine. Therefore, for any physical system we define this ability to do work as its **energy**. You should think of the energy of a system as a property of the system. Whenever a system does work on another system, this represents a transfer of energy from one to the other. The energy level of one decreases while the energy level of the other increases.

It is sometimes overlooked that in the calculation of work, time is not a factor. It does not matter whether the displacement involved takes one second or one year, the physical work is identical. If we want to take the duration of the motion into account, we need a new quantity. We define **power** to be the rate at which a machine or an engine performs work. The SI unit of power is called the **watt**.

It can also be said that the work performed by a system is the amount of energy produced by it (and power is the rate of its production). Approaching the definition of energy this way gives it a priority over the notion of work. We will see that this subtle change in perspective offers a completely different way of approaching mechanics—a way that does not use the notion of force.

When a force is applied to a particle otherwise free, the particle increases in speed. The work done by a constant force to accelerate a particle up to a given final speed from can be derived from equation (3.6) and Newton's second law (5.1). If we multiply the first by the mass m and the second by the displacement x we can combine them to yield

$$W = Fd = mad = \tfrac{1}{2}mv^2$$

This work done by the force increases the energy of the particle. Since this energy is manifest in motion we give it the name **kinetic energy**.[3] This means that any particle in motion has a certain amount of energy by virtue of this motion. Thus, we have derived a formula for this kinetic energy:[4]

$$KE_{\text{lin}} = \tfrac{1}{2}mv^2 \tag{7.2}$$

When we explore the theory of relativity in Lecture 26 we will see a need to revise this formula for particle speeds approaching the speed of light.

Work done on a system can manifest itself in a different way too. If a system is in stable equilibrium, its internal forces tend to pull it back to its equilibrium state. We do work against these internal forces when we displace the system away from this equilibrium. When a system is pushed out of stable equilibrium in this way we say that it has **potential energy**. The system has the ability to do work

[2] We have already used the symbol W for weight but we will now use it for work. This is a potential source of confusion. When they both appear in a problem, I will substitute mg for the weight and use W for work.

[3] This is sometimes given the symbol T.

[4] The subscript "lin" stands for linear. In Lecture 10 we will see in kinetic energy due to rotation. Until then we will generally drop the subscript.

because when released its motion back to equilibrium can be leveraged to perform other useful work.

One simple but important example of creating potential energy is lifting a weight. In order to raise an object against the pull of gravity, a force must be provided that pushes up. This lifting force must do work against the force of weight. Its magnitude is mg so the work done by the lifting force is mgh where h is the height of the lift. Therefore, we define the potential energy[5] of an elevated weight to be

$$PE_{\text{wgt}} = mgh \tag{7.3}$$

In general, the forces in a system are not constant. In these cases, we continue to define the potential energy as the amount of work required to move the system from equilibrium to the particular state of interest. In general, this calculation will involve some integral calculus.

Suppose we release the system and allow it to return to equilibrium. The work done by the system will be at the expense of its potential energy as it reverses the displacements required to create the non-equilibrium state. This work creates the kinetic energy involved in the motion of the system. In other words, the internal forces of equilibrium transform the energy from potential into kinetic—the total sum of kinetic and potential energy is always the same. This is the **conservation of energy**. Note that due to the way potential energy is defined, the total energy of an isolated system is conserved by definition.

Reconsider the example of the block sliding down an inclined plane in Figure 5.2 from page 30. Let us use energy considerations to determine the final speed.

In the initial state, the larger mass is elevated 0.5 meters. We know this because it slides done the slope one meter inclined at 30°.

$$PE = mgh = (10)(9.8)(0.5) = 49$$

Since the system starts at rest, this is the only element of energy that is non-zero so the total initial energy of the system is 49 joules.

Now consider the final state of the system. The larger mass has released all of its potential energy. It still has some kinetic energy due to its motion, which we need to determine. So we have

$$KE = \tfrac{1}{2}mv^2 = (5)(v)^2$$

In addition, the smaller mass rises one meter, so it has some potential energy.

$$PE = mgh = (3)(9.8)(1) = 29.4$$

This mass is also moving with the same velocity as the larger mass. So it also has some kinetic energy:

$$KE = \tfrac{1}{2}mv^2 = (1.5)(v)^2$$

This means that the final energy of the system is given by

$$E = 29.4 + (6.5)(v)^2$$

Since the total energy is conserved, we know that the energy level of the final state is still 49 joules. Setting our equation for the final energy of the system to 49 allows us to solve for v. We get an answer of $v = 1.74$ just as in the previous analysis.

Clearly this approach using energy is much simpler than the previous approach using force. This is a perfect example of the advantages in using energy. The disadvantage is that we cannot answer every mechanical question this way. For

[5] This is sometimes given the symbol U.

example, we do not know the direction of the velocity. Certainly we know from simply looking at the diagram, but not from the energy calculations. In addition, questions involving time and duration require going back to the analysis with force. But questions about displacement and speed are usually able to be answered in this way.

So, work against inertia creates kinetic energy. Work against stable equilibrium creates potential energy. What about friction? Work against friction produces **heat**. We will study heat as a form of energy in Lectures 15–17. Although it is not wrong to say that this is a third form of energy, it is should be emphasized that this form of energy is not interchangeable with the other two mechanical forms of energy. We will see that it is possible to recover some of this heat energy but for now we will simply consider it lost. We say that friction is a **non-conservative force** because it destroys rather than conserves mechanical energy.

Because the work that is done by a non-conservative force is lost there is no potential energy associated with it. The way to determine whether a force is conservative or not is by calculating the work done on a round trip. If the total work done is zero, energy is conserved because all the work consumed by the system is eventually recovered.

A corollary is that the work done to move a conservative system between two states is independent of the path taken. If the work were different we could combine the two paths to make a round trip with non-zero work. This is another way to determine if a force is conservative: the work depends only on the initial and final states of the system and not on what happens in between.

In summary, external forces and non-conservative internal forces change the total energy level of a system. Conservative internal forces convert the energy in the system from one form into another. So if we ignore friction, we can say that energy is conserved in any isolated system.

Suppose we choose a particular state of the system as a reference point. The choice is arbitrary, but you will see that it makes sense to pick a point of stable equilibrium. If the forces in the system are conservative, when we calculate the work done to move to another state we know that this is equal to the potential energy of that state and that the number is unique—it does not depend on how we get from the reference state to the other state.

Formally we say that the potential energy is a function of the state of the system. In the simplest case we have a particle under the influence of a conservative force which depends only upon its position. Examples include the elastic force of a spring and the long-range force of gravity. For the force of gravity, the potential energy function is[6]

$$PE_{\text{grav}} = -\frac{GMm}{r} \tag{7.4}$$

A much simpler, almost trivial, example is the force of weight. Since the force is constant, the potential energy function is simply mgh. On a graph this is a straight line with a slope of mg, which is just the weight again. In general, we can recover the force from the potential energy by taking the derivative of the potential energy function:

$$F = -\frac{dU}{dx}$$

The negative sign is there because a system will always tend to release its potential energy unless some obstacle stands in the way.[7]

Consider the potential energy plot in Figure 7.2. The white dot represents a point of stable equilibrium and the × marks a point of unstable equilibrium. Remember that force points against the slope, so the system is always pushed in the direction that reduces its potential energy. At the point of stable equilibrium, the forces point in while the force point out at the point of unstable equilibrium.

[6] Gravity is a bit of a special case because there is no natural equilibrium point to use as a reference. Instead we choose to use the point where the interaction is zero—which is infinitely far away. Because gravity is attractive it always takes work to get away. This is why the gravitational potential energy is always negative.

[7] If you rearrange this formula as $F\ dx = -dU$ it is easier to see how this formula is a corollary of the definition of potential energy.

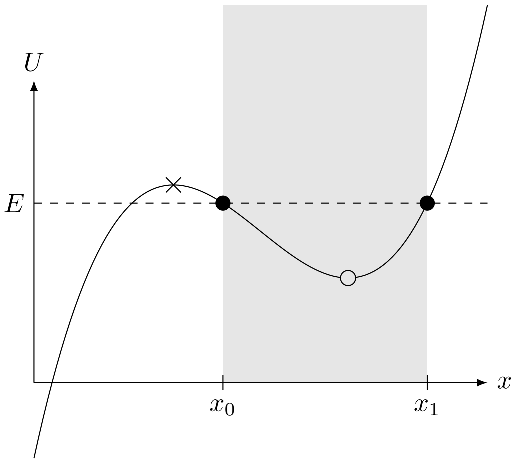


Figure 7.2: An example of an energy diagram. The white circle and × mark points of stable and unstable equilibrium, respectively. The black dots represent turning points where the motion of the system reverses. Outside the shaded area is "forbidden" to the system because the kinetic energy cannot be negative.

It is tempting to think of the energy diagram like a ball on a slope with the ball rolling down into the potential well. But the vertical axis represents the energy of the system. In this picture, the energy level is represented by the dashed line. If the energy is conserved, it is better to think of the system as shuttling back and forth on this dotted line.

Notice the black dots in the diagram. These are called **turning points**. Why? Remember that the graph represents the potential energy of the system. Since the dotted line represents the total energy, the gap in between the two must represent the kinetic energy. However, unlike potential energy, the kinetic energy can never be negative. This implies that the motion of the particle is constrained to be within the shaded region. The set of states for which the potential energy function exceeds the total energy of the system are "forbidden", and these areas are called the **forbidden regions** for the motion of the system. On the boundary of this forbidden region the potential energy equals the total energy and the kinetic energy equals zero.

What happens is this. Suppose the particle is moving toward a turning point. The force is opposing the motion because the slope is up in this direction. The kinetic energy of the particle is becoming smaller and smaller—it is slowing down. It does so until it reaches the turning point then, for a moment, it comes to rest. The force pushes it back and turns the motion around. This repeats over and over with the particle bouncing between the two turning points. In Lecture 12, we will learn that small displacements about stable equilibrium always exhibit simple harmonic motion.

We can take this one step further by incorporating a bit of friction. The effect of friction is to reduce the overall energy level. In other words, the dotted line will get lower over time. This has the effect of pulling the turning points closer to equilibrium until eventually the energy level settles down to the point of stable equilibrium. Thus, in the real world where friction is present, systems will always settle down to rest into the closest state of stable equilibrium. Once there they remain unless they are knocked out by some external force.

The principle of least action is a statement that physical systems move in such a way that they minimizes a certain property. Both Newton's first law and the principle of virtual work can be summarized by minimal principles. If we set the potential energy to zero, we are describing a force-free particle. The average kinetic energy multiplied by the total time is minimized by the shortest path—which is produced by a particle moving with constant velocity. In addition, if the kinetic energy is set to zero, we are describing a static system. In that case, the potential energy is minimized because the system is sitting at equilibrium in a potential well.

Since both the potential and kinetic energies obey a minimal principle, it is natural to suspect the total energy to do so also. But consider the projectile range problem. What prevents the particle from simply moving horizontally? Energy is conserved this way. Perhaps the particle moves along a path where is it minimized? No, the energy is actually minimized for a path that drops underground. Lagrange found the answer: subtract the potential energy rather than add it to the kinetic energy. This combination now bears his name: the **Lagrangian**. The average Lagrangian multiplied by the total duration is called **action** and depends upon the path.

$$L = KE - PE \quad \text{and} \quad S = \sum_{\text{path}} L\,\Delta t$$

The path with the smallest action is the true path. This is called the **principle of least action**.

We can at least indicate how to prove that Newton's second law and the principle of least action are equivalent. The simplest non-trivial example of using action is the step-potential. On either side of the potential we have a free-particle, so we expect the path to be straight. The only question is when does the particle encounter the step? We expect the particle to spend more time in the higher potential since this reduces the action. On the other hand, if the particle moves too quickly in the low potential region the kinetic energy will be too high. The optimal path is refracted (see Figure 7.3) from the constant velocity path. It can be shown that this path conserves energy, which we know comes from Newton's laws.

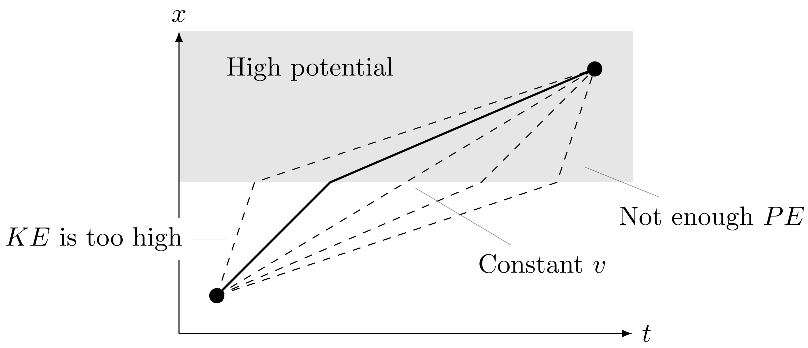


Figure 7.3: Refraction of a particle off of a potential step. Compare with optical refraction in Figure 20.7.

So the principle of least action not only fixes the energy, but also isolates the path. This principle is therefore more powerful than the energy approach. But that's only the beginning. This principle has applications to optics, wave motion, electromagnetism, general relativity, and even quantum mechanics.

For next week we will define momentum, a concept we mentioned in the discussion of Newton's second law back in Lecture 5 as the "quantity of motion". This will lead into a discussion about collisions and how to classify and analyze them using the idea of the center of mass. We will talk a bit about how to deal with multiple particles interacting. Finally we will get a sneak-peak at what quantum field theory looks like by discussing exchange particles and quasi-particles.

Lecture 8

Momentum and Collisions

Read sections 7.1–7.5, sneak a peak at section 14.3, review Lecture 5

In the 17th century there was quite a vigorous discussion about how to quantify motion. The debate ran across national lines with Newton in England and Leibniz in France.[1] We have seen in Lecture 5 that Newton based his work on the product of mass and velocity, which today we call **momentum**:[2]

$$p = mv \tag{8.1}$$

It combines the idea of inertia and motion, and because it is a vector also captures the direction of motion.

Leibniz, on the other hand, advocated the use of what he called ***vis viva***. Primarily based on observations of collisions, Leibniz noted that the product of mass and velocity squared is conserved under certain conditions. We can see now that this is twice the kinetic energy that we introduced in Lecture 7. So, both traditions are alive and well.

Momentum is directly related to the notion of force through Newton's second law. This essentially states that the rate at which momentum changes is equal to the net force on the object. We have also noted how Newton's third law can be interpreted as the statement that every interaction involves an exchange in momentum. One could also argue that Newton's first law can also be rewritten as a statement about momentum: in an inertial frame with no external forces, momentum is constant.

So, the idea of momentum is threaded throughout mechanics. We can take this one step further by rewriting the second law as:

$$F\,\Delta t = \Delta p \tag{8.2}$$

There are two things to say about this equation. The first is that the quantity on the left is called **impulse**. Impulse is usually only used for fast-acting forces, like contact forces in a collision. In this formula, the F represents the average force applied over the time frame.

The second thing is that this formula is similar to the formula between work and energy:

$$F\,\Delta x = \Delta E$$

In other words, every increment of work corresponds to an increase (or decrease) in the energy level of the system. Because of this similarity, the problem solving methods in this lecture are similar to those of the previous lecture. We identify the initial and final states of the system and the path in between is ignored. This is the great advantage of using the energy-momentum perspective: we don't need

[1] These two also had a priority dispute over who invented calculus. We still have competing notations that go back to these two men.

[2] Don't ask me why the symbol for momentum is p. I've never been able to figure it out.

to know the details of the interaction. We can't answer every question but the solution is much easier for those we can.[3]

Imagine a bullet embeds itself into a wall. The initial momentum of the bullet suddenly becomes zero. This happens because the wall provides the force necessary to stop this momentum. Now imagine a machine gun which fires 4.0 gram bullets at 940 meters per second at a rate of 720 rounds per minute. What is the average force per second required to stop these bullets?

It's easy to see that the average momentum of each bullet is[4]

$$p = (0.004)(940) = 3.76$$

So the wall provides the force against this momentum for each bullet. But in order to use equation (8.2) we need a time frame. What do we use in this case? We do not know how long it takes for the wall to stop each bullet. But we are not asked for the force to stop a bullet. We are asked for the average force. If we were to plot the force from the wall over time we would just see a bunch of spikes (see Figure 8.1). The average level of force is not the height of these spikes but somewhere below depending on how far apart the spikes are and how thick they are.

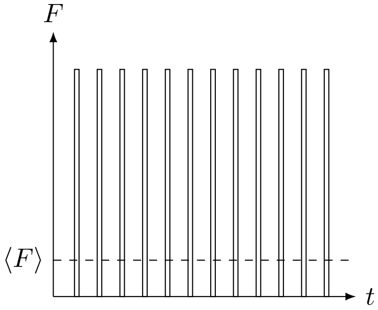


Figure 8.1: Average force required to stop the bullets

Fortunately these things all average out over time. What we need to do is just pick a time frame—say one second. How many bullets strike the wall in one second? The answer is 12. So the wall must provide the momentum to stop 12 bullets, or

$$p_{\text{tot}} = (12)(3.76) = 45.12$$

Since the wall does this in one second we can use equation (8.2) to solve for the average force. In this case the time frame is simply one, so the final answer is 45 newtons per second.

Though generally associated with electronics, the idea of a stream of particles occurs several times in physics and is called **current** (see Lecture 24). We will use the ideas from this problem again in Lecture 16 to derive the ideal gas law.

Momentum is also helpful in analyzing collisions. Consider a 2000 kilogram car moving at 20 meters per second that strikes another at rest. The second car has a mass of 2500 kilograms. How far they slide if they end up entangled together? Assume the coefficient of kinetic friction is 0.70.

Since the second car is at rest, it has no momentum. So the initial momentum of the system is just the momentum of the first car: 40,000 kilogram-meters per second. The collision does not change the overall momentum, so after the collision the momentum is the same. But now the momentum moves a total mass of 4500 kilograms. This implies that the speed after the collision will be 8.89 meters per second.

But this is only the first half of the problem. Now we need to know what it will take to bring this mass to rest. There are two ways to solve it: we can use force and friction from Lecture 5 or the ideas of work and energy from Lecture 7. Let's do it both ways. We will start with force.

The force of friction is given by μN where the support force is counterbalancing the weight of the mass. So,

$$F = (0.70)(4500)(9.8) = 30780$$

By Newton's second law this will cause the mass to decelerate:

$$a = \frac{F}{m} = \frac{(30780)}{(450)} = 6.86$$

[3] There is also another reason to notice the parallel between these two formulas. In Lecture 26 on special relativity we will see the concepts of time and space merge together into a single space-time continuum. At the same time we will see the ideas of energy and momentum merge together as well. This parallel is the reason for it.

[4] What is the SI unit for momentum? It is one of the few units in physics with no name. It one kilogram-meter-per-second. I usually just say "unit of momentum" rather than that mouth-full.

Since we know the velocities and we are interested in a distance, we should use equation (3.6) to answer the question. We have:

$$x = \frac{v^2}{2a} = 5.7589$$

Now let's turn to energy. The kinetic energy of the mass just after the collision is:

$$KE = \tfrac{1}{2}mv^2 = 1.7778 \times 10^5$$

This energy is consumed by the work done by the friction. This work is directly related to the distance, by definition. Thus,

$$x = \frac{W}{F} = \frac{(177780)}{(30780)} = 5.7757$$

So, the final answer is 5.8 meters.[5]

Not all collisions end with the objects stuck together: sometimes they bounce. There is a whole spectrum of collisions possible between the two extremes. On the one side you have an **elastic collision**. Suppose you take a ball and drop it from a certain height. If the ball returns to the same height the collision is said to be elastic. Of course it does not normally make it back. In that case we say the collision is inelastic. So an elastic collision is an ideal: the typical case is inelastic. If the ball drops and stops (like a sloppy piece of mud), we call this a **completely inelastic collision**.

This spectrum is quantified by the **coefficient of restitution** (COR) which is related to the speed prior to and after the impact.[6]. An elastic collision will have a COR equal to one and a completely inelastic collision will have a COR equal to zero.

The important thing to remember is that the momentum in any collision, no matter how inelastic, is always conserved. Kinetic energy is what is lost in an inelastic collision, not momentum.

Based on this insight, we can derive a couple of useful formulas relating the initial and final velocities of a collision. In general, the mathematics is pretty hairy, so we will confine ourselves to problems in which one of the objects is initially at rest. This is not as restrictive as it may sound. According to the principle of relativity, we have the freedom to choose our inertial frames. We can choose the frame that follows the particle, effectively making it at rest. This is called the **lab frame** for the collision.

A couple of conventions first. It is usual to label the first particle as the one moving and the second particle as the one at rest. In order to not get lost in a sea of subscripts I prefer to use different letters to represent the initial and final velocities. This is not usual but neither is it unheard of in the literature. We will use v to represent the initial velocities and u the final velocities. So, what we seek is a formula for u_1 and u_2 given the fact that $v_2 = 0$.

For every collision we have the conservation of momentum.

$$m_1 v_1 = m_1 u_1 + m_2 u_2$$

Since we are looking for two unknowns we need another (assume the masses are given). In general this involves the coefficient of restitution, but for the extreme cases we have a simpler fall-back. For a completely inelastic collision we have $u_1 = u_2$ since they stick and travel together. This allows us to derive:

$$u_1 = u_2 = \left(\frac{m_1}{m_1 + m_2}\right) v_1 \tag{8.3}$$

For a (perfectly) elastic collision the total kinetic energy is conserved. This follows from the fact that the COR is equal to one.[7] If we multiply the kinetic energy

[5] The difference between these answers is a rounding error—which is why we need to pay attention to significant digits!

[6] The bounce height is related to the (potential) energy of the ball, so the COR is equal to $\sqrt{h/h_0}$.

[7] We will be able to establish this after we have discussed the center of mass.

equality by two we get

$$m_1 v_1^2 = m_1 u_1^2 + m_2 u_2^2$$

You can see here the origin of Leibniz' *vis viva*. The math is a bit messy, but after a while we get

$$\begin{aligned} u_1 &= \left(\frac{m_1 - m_2}{m_1 + m_2}\right) v_1 \\ u_2 &= \left(\frac{2m_1}{m_1 + m_2}\right) v_1 \end{aligned} \tag{8.4}$$

Notice that if $m_1 < m_2$, the incoming particle will bounce backward. If $m_1 \ll m_2$, the particle will bounce back with nearly the same speed and the struck particle will hardly move—like a wall.

Another example of the use of momentum is rocketry. People often ask: in space, how can a rocket move—what does it push against? The answer is that it pushes against its own exhaust. Consider a kid sitting in a wagon at rest. Suppose she takes a baseball and throws it. She and the wagon recoil slightly. This is the same recoil which drives a rocket. The momentum calculation is the same as a completely inelastic collision, just in reverse. In fact, we can even incorporate this explosive type of interaction on our COR spectrum if we let the numbers exceed one because an explosion is an increase in the total kinetic energy of the system.

The complication with rocketry is that it requires a lot of fuel to push the rocket. This means that it is not reasonable to assume the mass of the rocket is constant. So, at each instant the momentum calculation will change. If one does the math,[8] one gets the rocket equation which describes the overall increase in speed.

$$\frac{\Delta v}{v_e} = \ln\left(1 + \frac{\Delta m}{m}\right) \tag{8.5}$$

where v_e is the speed of the exhaust, Δm is the total mass of the fuel, and m is the final mass (without the fuel). This **delta-v** is important in calculating orbital maneuvers in space. In general, the mass of fuel required far outweighs the mass of the payload.

So far we have only spoken of these collisions in one dimension. Consider again a collision with an incoming particle colliding with another at rest. Imagine the approach of the moving particle to be off-center like a billiards shot (see Figure 8.2). The contact forces line up with the point of contact so they are at an angle. This deflects the first ball and sends the second down the line creating a truly two-dimensional problem.

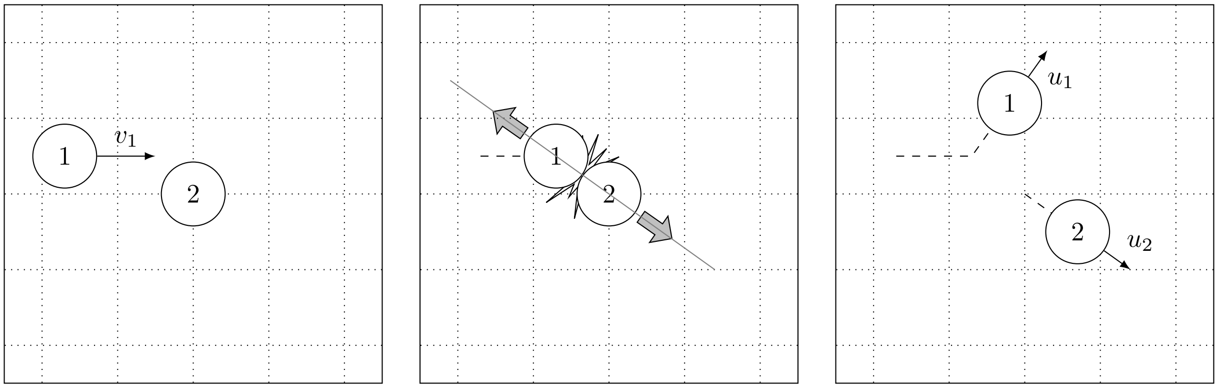


Figure 8.2: Two-dimensional collision (lab frame)

In any scattering problem like this, the final angle of deflection depends upon the nature of the interaction between the particles. Particle accelerators probe the

[8] This is an example of an application of Newton's second law with variable mass.

inner working of the nucleus by analyzing deflection patterns in order to infer the nature of the interaction.

Deflection will even occur for an attractive force. This is how a **gravitational sling-shot** works. Passing by a planet, a space probe can steal momentum from the planet and be deflected along a new trajectory. This is one way to pick up delta-v without expending fuel.

For any collision problem, it is usually easiest to perform the analysis in the **center of mass frame** (or CM frame). This is the inertial frame in which the system's center of mass is at rest. The **center of mass** is defined as

$$x_{\text{cm}} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$$

where x_1 and x_2 are the positions of objects.[9] It can be seen that the velocity of the center of mass is constant because we have by definition:

$$v_{\text{cm}} = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}$$

The numerator is the total momentum of the system and the denominator is the total mass of the system. In any collision, elastic or not, both of these quantities are conserved.

The reason the CM frame is helpful is that the velocity of the CM is zero so the total momentum is zero. This implies that the momenta of the particles are equal and opposite before and after (see Figure 8.3).

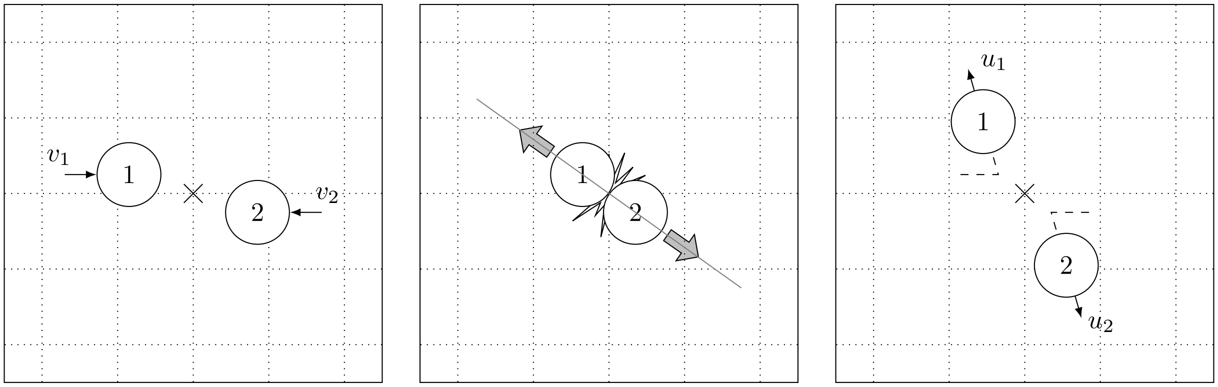


Figure 8.3: Two-dimensional collision (CM frame)

So far we have only considered two particles interacting. What if there are more? The definition of the center of mass extends to include all the particles. It's still true that the momentum of the total system is conserved (and the total mass), so the velocity of the center of mass is constant. In the CM frame, all the velocities are measured relative to the center of mass so the aggregate whole does not move. But the parts move all over. Internal forces are responsible for redistributing the energy and momentum of the system but they cannot change the motion of the whole.

We will see in Lecture 16 that the total kinetic energy of the molecules in the center of mass frame for a macroscopic system is directly related to its temperature.

Quantum field theory (QFT) describes the fundamental interactions of nature as occurring through the exchange of intermediate particles. These particles exchange momentum, energy, and other subatomic qualities. Figure 8.4 shows a **Feynman diagram** used in a QFT calculation. At each vertex momentum, energy, and all the other subatomic qualities are conserved. See Lectures 27 and 30 for more details.

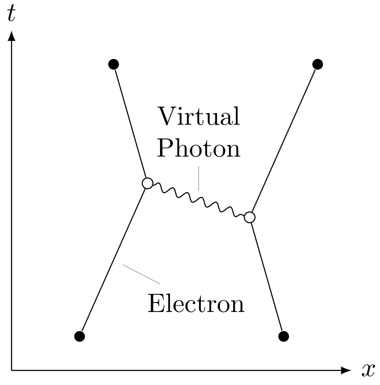


Figure 8.4: A simple Feynman diagram

[9] Sometimes we are only told the distance between the two objects. In that case, let x_1 equal zero, and let x_2 be the distance given. The center of mass will then be given as a distance from the first object.

Next week we will expand our mechanics for the first time beyond the particle and consider the rotation of a rigid object. We will see an immediate parallel with the kinematics in Lecture 3. This parallel will tee up an advanced topic called generalized coordinates. The idea of rotation will be seen as a special example of generalized coordinates and we will talk about rotating coordinate systems.

Lecture 9

Rotation and Non-Inertial Frames

Read sections 8.1–8.6

In Lecture 4 we discussed the way in which force can affect an object. We agreed to focus on the motion of a particle and systems of particles. In this lecture we will broaden that focus to include objects with extension. We will still ignore the possibility of deformation (see Lecture 13). An object in which the deformation is negligible is called a **rigid object**.

A rigid object may be composed of an arbitrary number of parts. But those parts are constrained to move in such a way that their relative distances are unchanged in time. For any system, the various ways in which it can move are called its **degrees of freedom**. A single particle has three for the three dimensions of space. A collection of N independent particles will have $3N$ degrees of freedom. A rigid object has six: three for its bulk motion in space and three for its rotation about the center of mass.

In general, when an object is tumbling through space it typically has a wobble in its rotation. This **free rotation** is due to some lack of symmetry in the object. We will touch on ways to analyze this general case in Lecture 10, but for now we will start simple. We will assume that the axis of rotation is fixed so that the motion the parts are in parallel circles centered on the axis. This is called **fixed rotation** and will be the 90% of what we will consider in this class.

Our previous considerations of the motion of the center of mass are valid for rigid objects. For example, in the absence of external forces the center of mass of the rigid object will move with constant velocity so the CM frame is inertial. Effectively this allows us to ignore the bulk motion of the object and focus our attention on just its rotation.

Since the parts all rotate together, it is sufficient to watch the motion of one point—the rest will follow. We already mentioned that the motion of this point will be in a circle with its center along the axis of rotation. So everything you know about circles is relevant. We will review the concept of arc-length.

The **arc-length** is the curved distance along the circumference of the circle between two points on the circle. There is a one-to-one correspondence between the angle formed between these points from the center and this arc-length. It is easier to characterize the rotation of the object by this angle since the parts form a consistent angle as the object rotates, but the arc-lengths may be quite different.

However, the angle is not a distance—the arc-length is. And mechanics is mostly about tracking displacements over time. So we need to be able to convert between the two ideas. The arc-length is clearly proportional to both the angle and the radius of the circle. Also, if the angle is 360°, we know the arc-length is the full circumference of the circle, $2\pi r$. So the formula is

$$s = (2\pi r)(\theta/360^\circ)$$

We can simplify this formula by introducing a new unit to measure angles called the **radian**. If we define the radian such that 2π radians equals 360 degrees, the formula becomes

$$s = r\theta \tag{9.1}$$

where θ is now measured in radians. This means that an angle of one radian will produce an arc-length equal to the radius of the circle. This new unit is just a different way to measure angles—like the difference between centimeters and inches. One radian is a smidge less than 60°. The radian also simplifies other calculations involving angles: for example, in Lecture 12 we will use the fact that $\sin\theta \approx \theta$, a trick which only works with radians.

Another common unit for measuring angles is the revolution—as in rpm, revolutions per minute. The conversion factors between the three units of angle are simply

$$1 \text{ rev} = 2\pi \text{ rad} = 360^\circ \tag{9.2}$$

If we consider our arc-length to have a direction (i.e., from point A to point B), it starts to sound like a vector. Technically it is not because vector addition will not work with arc-lengths,[1] but we do want to keep the idea of "directionality" for our analysis of motion. So we will avoid calling these vectors, but we will still want to use the idea of a directed arc-length.

In addition, we want to associate a direction to our angles. This is called **angular displacement**. Remember, when we talk about angular displacement we are talking about an angle: its unit is either degrees or radians.

This leads into the idea of angular velocity which is simply the rate at which the angular displacement changes. Its units are radians per second. Note that revolutions per minute is a unit of angular velocity. We will also speak of angular acceleration, which is the rate at which the angular velocity changes. The traditional symbols for angular velocity and angular acceleration are ω and α, respectively.

Now take a minute and review the logic we used to derive equation (3.1)–(3.6) in Lecture 3. They all follow from the definition of velocity and acceleration. The definitions of angular velocity and acceleration are built on a parallel between linear displacement and angular displacement from equation (9.1). So we can import all these linear equations into the context of angular rotation.

Let's consider an example. A wheel is spinning at 30 revolutions per minute. What angular deceleration is required to bring the wheel to rest in five seconds?

We are told the initial speed, though we do want to convert this into radians per second.

$$\omega_0 = \frac{30 \text{ rev}}{\text{min}} \times \frac{2\pi \text{ rad}}{1 \text{ rev}} \times \frac{1 \text{ min}}{60 \text{ s}} = 3.1416$$

The final speed is zero since it comes to rest. Since we are also given the time involved we should use the angular equation that corresponds to equation (3.2). That is,

$$\omega = \omega_0 + \alpha t$$

and

$$(0) = (3.1416) + (\alpha)(5) \implies \alpha = -0.63$$

Any other problem like this can be solved using the methods from Lecture 3.

[1] It kind of works in two dimensions, but it breaks down for free rotation. It is possible to construct a vector space, but only for infinitesimally small arc-lengths and rotations

How are these ideas related to those in Lecture 6 on circular motion? In that lecture, the speed was associated with the motion along the circumference or the arc-length. It's pretty straight-forward to see that this is related to the angular velocity by

$$v = r\omega \tag{9.3}$$

which follows from equation (9.1). Similarly we have

$$a = r\alpha \tag{9.4}$$

for the angular acceleration. But be careful: this acceleration is the acceleration along the rotation. The acceleration calculated in equation (6.1) is directed toward the center. These two vectors are perpendicular to one another.

In order to avoid confusion we will speak of the components along the circular motion as tangential and the components toward the center as centripetal. For example, we add t as a subscript and now describe equation (9.3) by saying that when the angular velocity is constant, the tangential velocity is proportional to the distance from the axis of rotation.

It is worth noting that the polar coordinates we introduced on page 35 are consistent with this division of motion into tangential and centripetal.

One way these concepts are used is in rolling motion. The condition for a rolling wheel is that its rotation match its translation. If this is not the case, the wheel is slipping (or sliding if it is not rotating at all). In symbols, we require

$$v_{\text{cm}} = v_t = r\omega \tag{9.5}$$

Consider the following problem. Suppose a go-cart with wheels of radius 0.1 meter has an engine that can produce an angular acceleration of 2.0 radians per second squared. What is its final speed at the end of a 100 meter race? Assume the acceleration is constant.

Since the problem involves acceleration, speeds, and distance, we will use equation (3.6). In terms of angular variables we have

$$\omega^2 = \omega_0^2 + 2\alpha\theta$$

The initial speed is zero and the final speed we will get from the final angular velocity. But what is θ? This is the total angular displacement of the wheel. Knowing that the wheel does not slip allows us to identify the 100 meter distance with the total arc-length travelled by the edge of the wheel. So, we can use equation (9.1) to calculate the angular displacement:

$$(100) = (0.1)(\theta) \implies \theta = 1000$$

Now we can solve for the final angular speed:

$$\omega^2 = (0)^2 + (2)(2.0)(1000) \implies \omega = 63.246$$

Since $v_{\text{cm}} = r\omega$, the final speed is 6.3 meters per second.

There is another way to solve this problem. Since the wheel is rotating, we can also say that $a_{\text{cm}} = a_t = r\alpha$. In this case $a_t = 0.2$ meters per second squared. Then we can use (3.6) directly to get the same answer.

This covers the basic kinematics of rolling and rotation. We will discuss dynamics (torque as the cause of rotation) in the next lecture. As a segue to discuss generalized coordinates I would like to talk about rotating reference frames.

A question may have already occurred to you from Lecture 5. How can we get away with considering our laboratories as inertial frames? We all know the earth is rotating and taking the lab with it. The short answer is that, technically, we cannot. Our labs are not inertial. But the effects are slight enough that we can neglect them.

Imagine you are located directly over a spinning merry-go-round looking down from the top. Every object on the merry-go-round has a tangential velocity given by equation (9.3). Suppose a child is playing with a ball and lets it go. From your viewpoint (the inertial one), the ball is free and follows a straight-line inertial trajectory. This is away from the center of rotation. But the child is held in place by the merry-go-round. He sees his ball traveling out and away. He feels the force of the merry-go-round obstructing his inertial motion. He is experiencing **centrifugal force**.

Do we know the formula for this inertial force? It is caused by the centripetal forces holding the merry-go-round frame together. Therefore it must also have the form given by (6.1). In terms of the angular speed of the frame, we have

$$F_{\text{cfg}} = mr\omega^2 \tag{9.6}$$

The **Coriolis force** is another, more subtle, inertial force associated with rotating frames. It is a bit difficult to derive, so I will simply quote the result:

$$F_{\text{cor}} = 2mv_t\omega$$

Notice that this is a force that requires the object to be moving before it is felt. Because the frame is moving underneath you an extra force is required to keep a moving object in line with the frame. In addition, the direction of the force is perpendicular to the velocity.[2] So the Coriolis force is a force of deflection—again, this is an artifact of the moving frame. In reality the particle moves in a straight line. In fact, the characteristic spiral motion of hurricanes is a result of the Coriolis force in the atmosphere.

So how big are these forces? Well, the earth rotates once per day, so its angular velocity is

$$\omega = \frac{2\pi \text{ rad}}{86400 \text{ sec}} = 7.27 \times 10^{-5}$$

The radius of the earth is 6.38×10^6 meters, so the centrifugal force on an object will be

$$\begin{aligned} F_{\text{cfg}} &= (m)(6.38 \times 10^6)(7.27 \times 10^{-5})^2 \\ &= (m)(0.0337) \end{aligned}$$

Compare this with the formula for weight, mg. The centrifugal force on an object is 0.34% of its weight. The Coriolis force is much smaller which is why it only manifests itself on the large scale.

Every reference frame (even a rotating one) is at rest relative to itself, by definition. No matter how wildly it moves, no matter how curvilinear (like polar coordinates)—the frame is the reference. Occasionally it is more convenient to follow the degrees of freedom in the system than to use an inertial frame. The downside is the introduction of inertial forces.

In the process of uncovering the principle of least action, Lagrange also investigated the use of an arbitrary non-inertial frame and he analyzed it's impact on Newton's laws of motion. His approach is sometimes called **Lagrangian mechanics**.

Essentially, the first step is to identify the degrees of freedom of the system. The coordinates are then chosen to align with the natural constraints in the system. Relative to these coordinates the configuration is measured. These are called **generalized coordinates** and are usually symbolized with the letter q. Angular

[2] Here is an example where the vector product is handy. We can write this as

$$F_{\text{cor}} = -2m\vec{v} \times \vec{\omega}$$

with the angular velocity vector pointing along the axis of rotation.

displacement is an example of a generalized coordinate. The definitions of velocity and acceleration are generalized also. They are usually symbolized by $\dot{q}$ and $\ddot{q}$, respectively.[3]

Building on the principle of least action, Lagrange was able to generalize Newton's second law into a form appropriate in any coordinate system, inertial or not. These equations (called the **Euler-Lagrange equations**) of motion naturally incorporate the inertial forces in the frame. In addition to all the other uses of generalized coordinates, they are helpful in the study of general relativity where space-time is curved. The identification of gravity as an inertial force (the equivalence principle, see page 154) follows directly in this approach.

Next week we will see the analogy between linear displacement and angular displacement expanded. The notion of torque will be defined and we will talk about equilibrium in rigid systems. Newton's second law has an angular version which will require us to define a type of inertia for rotation. We will use this to define angular momentum and kinetic energy. We will finish off by touching on free rotation and gyroscopic motion.

[3] It's a little calculus convention to symbolize the rate of change of a quantity with a dot. This notation goes back to Newton.

Lecture 10

Torque and Free Rotation

Read sections 9.1–9.6, look back at section 8.7

Suppose a rigid object is constrained to rotate about a particular (fixed) axis of rotation. When a force is applied to this object, it will rotate. The amount of rotation is proportional to both the force applied and the distance from the axis. This is the principle behind the lever. This combination of force and distance is called **torque**.

If the object is free then the force will also push the object in a straight line based on Newton's second law (5.1). So the object will tumble in a combination of linear and angular motion. We can always break the analysis of this motion into the linear motion of the center of mass and the angular motion about the center of mass.

We can add an additional force to counter-balance the first. This combination will have a net force of zero, but it may still cause rotation. A balanced pair of forces with a non-zero torque is called a **force couple**.

There are actually two ways to calculate torque. One way is to draw a radial line from the axis of rotation to the point at which the force is applied. Now calculate the components of the force against that radial line. The torque is the radial distance multiplied by the tangential component of the force. See Figure 10.1. Though consistent with the previous idea of tangential and centripetal components, this is not the standard approach.

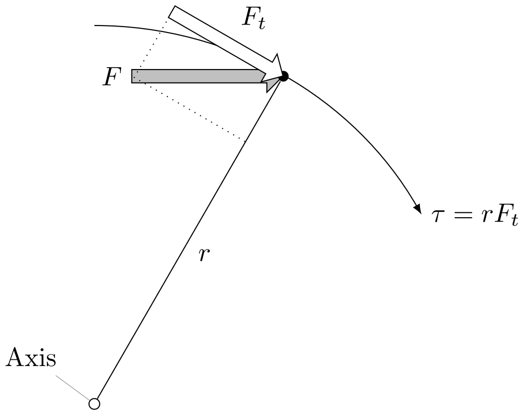


Figure 10.1: The tangential method of calculating torque

The standard way to calculate torque is through the idea of **lever arm**. The first thing to do is to draw the **line of action** through the force. The perpendicular distance between this line and the axis is the lever arm. See Figure 10.2. Increasing this distance will increase the leverage of the force.

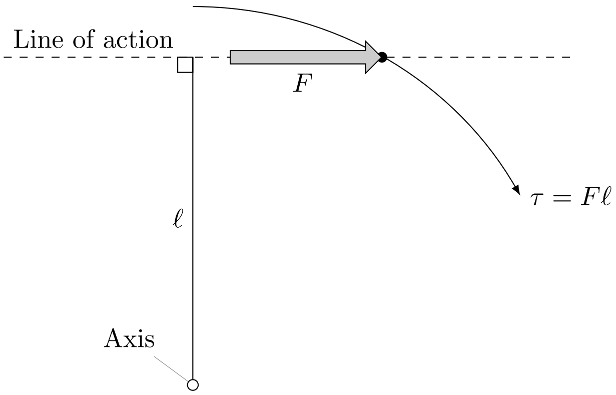


Figure 10.2: The lever arm method of calculating torque

For a rigid object to be in equilibrium all the forces and all the torques must balance. This is because there must be equilibrium along all the degrees of freedom of the system. For now we focus on rotation in a fixed plane, so three of the degrees of freedom are eliminated.[1] This leaves three degrees of freedom for our rigid body: two for linear motion and one more for rotation. Therefore, we can summarize equilibrium for a rigid body with the formulas:

$$\sum F_x = 0 \qquad \sum F_y = 0 \qquad \sum \tau = 0$$

As an example, consider a storage shelf which is held up by two posts separated by one meter. A two kilogram mass is sitting 0.4 meters from the left post. What is the support force provided by each post? See Figure 10.3. I think you can guess that the left post will hold 60% of the weight and the right will hold the remaining 40%. Let's confirm.

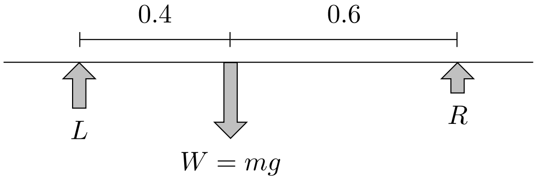


Figure 10.3: A supported shelf

First of all, there is no horizontal component to this question. All the forces are vertical, so $\sum F_y = 0$ is true but unhelpful. The formula $\sum F_x = 0$ tells us that the forces balance, so we have

$$L + R = mg = 19.6$$

The torques will give us another equation to solve this problem. In order to calculate the torque we need to measure the lever arm relative to the axis of rotation. But there is no axis of rotation: the shelf is in equilibrium! In this case, its arbitrary. Choose any axis you want. The reason is that the torques around any axis must balance—if they didn't, the object would spin there.

Though the choice is arbitrary, a judicious choice can make the math easier. Basically you can eliminate one torque calculation by putting the axis on top of the force you dislike the most. This example is pretty simple, but for more complicated problems you may want to think about this a bit. For now I will choose to put the axis right where the weight is applied.

In this problem the forces are all perpendicular to the lever arms so there is no trig to mess with. The torque from the left support is simply

$$\tau_L = (L)(0.4)$$

The right torque is similar:

$$\tau_R = (R)(0.6)$$

[1] This is because the fixed axis of rotation involves two additional constraints and the third is that the object is not sliding up and down this axis.

Before we add these together recognize that these two torques are in opposition. The left torque wants to rotate the shelf clockwise but the right torque wants to rotate the shelf counter-clockwise. The convention is to call counter-clockwise positive[2] so we should flip the sign of the left torque. The torque formula gives

$$-(L)(0.4) + (R)(0.6) = 0$$

So R is two-thirds the size of L. After substituting that into the first and solving we confirm that L is, in fact, 60% of the weight.

The previous problem spells out the general approach to solving these torque equilibrium problems:

- Set the sum of the horizontal and vertical force components to zero.
- Pick an axis of rotation with an eye to simplifying the calculation of the torques.
- Calculate the lever arm for each force (one of them should be zero).
- Sum the torques and don't forget to determine the correct sign to give them.
- Work the algebra—this usually involves solving multiple equations.

Let's do another problem that involves some torque calculations.

Consider a object that is hanging off of a wall. The object is held 2.0 meters off of the wall by a rigid support beam which weighs 50 newtons. The beam is held up by a string attached to the wall and makes a 30° angle with the beam. The maximum tension the string can support is 500 newtons. What is the largest weight that can hang off of the beam before the string snaps? Refer to the diagram in Figure 10.4.

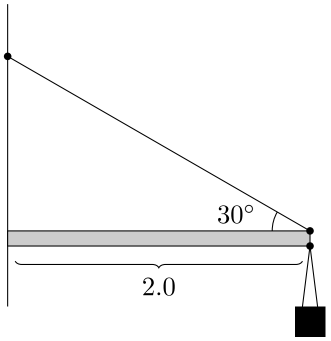


Figure 10.4: A hanging weight

Okay. There are four forces on the beam:

- The support from the wall which is probably angled up and out
- The tension in the string which is angled up and in at 30°
- The weight of the beam
- The weight of the hanging object

In terms of the horizontal, only the first two contribute. Since we want to maximize the tension in the string, we will consider its magnitude to be 500 newtons. Its horizontal component is

$$T_x = (500)(\cos 150°) = -433.01$$

So the horizontal component of the support is 433 newtons out.

For the vertical, we have all four forces in play. We are looking for the weight of the object, so we will label it W. The vertical component of the tension is

$$T_y = (500)(\sin 150°) = 250$$

We know the beam has a weight of 50 newtons down. The only force remaining is the vertical component of the support from the wall, which we don't know. Let's call it F for now. So, the equilibrium equation for the vertical forces is

$$-W + (250) + (-50) + (F) = 0$$

Since the weight points down I gave it a negative sign.

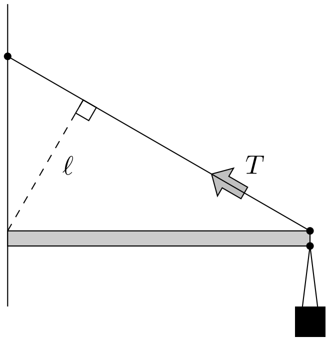


Figure 10.5: Calculating the lever arm of the tension in Figure 10.4

Now we move to the torques. The first step is to pick our axis. Since the support from the wall is unknown, let's put the axis there so its lever arm of the support

[2] This matches how we measure angles on our coordinate grid. The rule is that the angle is positive angle if it is measured from the positive x-direction toward the positive y-direction.

will be zero. The weight of the beam is concentrated at its center[3], so it has a lever arm of one meter. The hanging object has a lever arm of two meters. The last one to determine is the lever arm of the tension. The line of action is along the string, so the dotted line in Figure 10.5 represents the lever arm. This line is the opposite side of a right triangle whose hypotenuse is the length of the beam. The cosine connects the two. Therefore

$$\ell = (2)(\cos 30^\circ) = 1$$

Force	Magnitude	Lever Arm	Direction	Torque
Tension	500	1	+	500
Beam weight	50	1	−	−50
Object weight	W	2	−	$-2W$

Table 10.1: The torque calculations for Figure 10.4

Table 10.1 summarizes what we have so far. I have left off the support from the wall since we deliberately set its torque to zero. These torques have to add up to zero, so

$$(500) + (-50) + (-2W) = 0$$

Well here is a surprise. We can simply solve for W directly from the torque equation. The maximum weight the beam can hold is 225 newtons.

A couple of notes from this solution. Notice how we were able to avoid a bunch of math by picking our axis to zero out the support force. In fact, we never even needed to look at the horizontal and vertical components of the forces.

The second thing is that the calculation of the lever arm for the support was the worst part of the problem. This is usually the case. It always comes down to finding a right triangle and using our trig. If you have multiple lever arms to find, this can be quite annoying. Another approach is to calculate the torque tangentially. In this problem that means we need to determine the vertical component of the tension—which we calculated earlier as 250 newtons. Since the distance from the axis for this force was two meters, this also yields a torque of 500 newton-meters. You can see that this approach gets to the torque much faster.

Depending on the information given in your problem you might consider using one or the other approach. In the end they are equivalent, so it's up to you.

We are now ready to consider non-equilibrium problems in which the torques do not cancel. You may worry that this will be unbearably difficult given the complications involved in solving the corresponding torque equilibrium problems. However, we've done most of the heavy lifting already, so take heart.

Suppose we have a particle attached by a small rod to a particular point in space without any external forces. If in motion, it will move with uniform circular motion. This is because the centripetal force from the rod is always perpendicular to the motion, so the speed will not change. We can summarize this by saying that the angular acceleration of the particle is zero. This is analogous to Newton's first law of motion.

Now suppose we apply a force to the particle. Any component of the force perpendicular to its motion will be counter-balanced by the rod, so we only need consider a force parallel with the motion. This force will accelerate the particle according to Newton's second law, equation (5.1). We also know that torque is related to force via $\tau = Fr$ and acceleration to angular acceleration via $a = r\alpha$. So we can rewrite Newton's second law as

$$(\tau/r) = (m)(r\alpha)$$

[3] This point is called the **center of gravity**. It coincides with the center of mass unless the force involved varies over the configuration of the system. For standard gravity (i.e., weight) the force is constant so there is no difference.

or, more suggestively,

$$\tau = I\alpha \tag{10.1}$$

where $I = mr^2$ is called the **moment of inertia** for the particle. Equation (10.1) represents the angular analog of Newton's second law.

For an extended rigid object, this formula is the same. Essentially, the extended object can be considered as a (large) collection of particles, so the moment of inertia in general is

$$I = \sum mr^2 \tag{10.2}$$

If mass measures the tendency of an object to maintain its linear motion then this moment of inertia measures the tendency of an object to maintain its angular motion. The moment of inertia depends not only on the total mass but also on the distribution of that mass throughout the object.

For simple shapes, the formula for the moment of inertia can be written in terms of the total mass. For example, a hollow sphere rotating through its center has a moment of inertia of $\frac{2}{3}MR^2$. See Table 10.2 for a more complete list.

Ring	MR^2
Disk	$\frac{1}{2}MR^2$
Sphere	$\frac{2}{5}MR^2$
Shell	$\frac{2}{3}MR^2$
Rod (center)	$\frac{1}{12}ML^2$
Rod (edge)	$\frac{1}{3}ML^2$

Table 10.2: Moment of inertia for various systems. See here for more.

Consider the following example. A disk has a radius of 0.10 meters and a mass of 2.0 kilograms. It is attached on-center to a circular post (radius = 0.01 meters) which allows it to rotate horizontally (assume no friction). An ideal string is wound around the post and is attached to a 0.5-kilogram object that is hanging over a pulley. As this object falls, the disk rotates. After three seconds, how fast is the disk rotating? See Figure 10.6

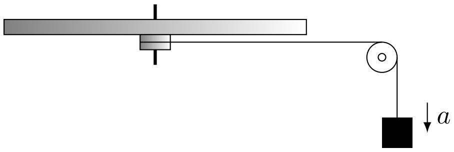


Figure 10.6: Rotating disk attached to a weight

The disk begins rotating because of the torque applied from the string. Since this is a non-equilibrium problem, we should not assume the tension in the string equals the weight of the hanging object—it will be less because the mass falls. In fact, Newton's second law tells us that

$$T - mg = -ma$$

where T is the tension in the string and m is the mass of the hanging object. This is just like in the inclined plane example on page 23. So we have a formula between the tension and the acceleration:

$$T = 9.8 - (0.5)(a)$$

Now consider the disk. For this we have Newton's second law in its angular form (10.1). The moment of inertia for a disk is $I = \frac{1}{2}MR^2$, so for this problem we have

$$I = \tfrac{1}{2}(2.0)(0.10)^2 = 0.020$$

The torque is from the tension in the string which it is applied to the post, so the lever arm is its radius:

$$\tau = (T)(0.01)$$

The angular acceleration of the disk can be gotten from equation (10.1):

$$(T)(0.01) = (0.020)(\alpha) \implies \alpha = (0.5)(T)$$

Using the previous expression for the tension yields

$$\alpha = 4.8990 - (0.25)(a)$$

Now, which radius do we use in $a = r\alpha$ to solve this equation? The acceleration of the block is equal to the tangential acceleration of the post because these are connected by the string. So, $a = (0.01)(\alpha)$. Since we are asked about the disk rather than the object, we should use this to solve for the the angular acceleration:

$$\alpha = 4.8990 - (0.02)(0.01)(\alpha)$$

So $\alpha = 4.8990$. In order to answer the question we should use the angular analog of equation 3.6:

$$\omega^2 = (0)^2 + (2)(4.8990)(3) \implies \omega = 5.4216$$

After three seconds the disk is rotating at 5.4 radians per second, or just under 52 rpm.

Shall we do another? Suppose a sphere rolls down an incline of 30° which is 4.0 meters long without slipping. How fast is it moving at the bottom of the incline?

For this one we will use energy. Since causing an object to rotate requires work, we expect a formula similar to equation (7.2) for the kinetic energy of rotation:

$$KE_{\text{rot}} = \tfrac{1}{2} I\omega^2$$

In this case, all the initial energy is potential. The sphere is at a height of 2.0 meters, so its potential energy is

$$PE = (m)(9.8)(2.0) = (19.6)(m)$$

At the bottom of the slope this potential energy is converted into kinetic, but the kinetic energy is split between linear motion and rotation:

$$(19.6)(m) = \tfrac{1}{2} mv^2 + \tfrac{1}{2} I\omega^2$$

We know the sphere does not slip, so we also have $v = r\omega$ and the moment of inertia for a sphere is $\frac{2}{5}MR^2$. Since we are asked for the speed we combine these to write

$$(19.6)(m) = \tfrac{1}{2} mv^2 + \tfrac{1}{2}(\tfrac{2}{5}mr^2)(v/r)^2$$

or,

$$19.6 = \tfrac{7}{10} v^2$$

The final speed will be 5.3 meters per second. This speed is independent of the mass and radius of the sphere.

In addition to kinetic energy we also have a rotational analog for momentum. The **angular momentum** of a system is

$$L = I\omega \tag{10.3}$$

which just like linear momentum is conserved in any collision (assuming no external torque). This conservation is why it is easier to ride a bike that is moving than balancing on one that is stationary. Once the angular momentum of the wheels is built up, they tend to continue rotating along the same axis.

That completes the analogy between linear motion and rotation. Except we have only scratched the surface. So far we have only discussed rotation in a plane with a fixed axis. We have yet to talk about free rotation. In general as an object tumbles it rotates with a wobble. In order to understand this we will need to refine our description of rotation.

Since each rotation involves a magnitude (the angular speed) and a direction (the axis of rotation), we can talk about it like a vector. For fixed rotation in a plane, this vector points out of the plane.[4] In addition, we can associate angular acceleration, torque, and angular momentum with vectors.

How then do we explain the wobble? It happens that, for any shape object, there are certain **principal axes** about which the object will rotate without this wobble. In this case the angular velocity vectors and the angular momentum vectors are aligned. For rotation off of these axes, the two are not aligned. The distribution of the mass throws off the balance of the rotation.

Since the distribution of mass is captured in the moment of inertia, the angular momentum captures both the rotation and this distribution. And since the

[4] Whether it point out or in is arbitrary. It is conventional to consider the vector as pointing at you when you are looking at counterclockwise rotation. This is consistent with the xyz axes in the sense that rotation from the positive x-direction to the positive y-direction points in the positive z-direction.

angular momentum is the conserved quantity (assuming no external forces), the angular velocity will have a tendency to rotate about the angular momentum. This twirling of the axis of rotation is called **gyroscopic precession**.

This is why you will sometimes see the moment of inertia mentioned as a tensor: it is the thing that connects the two vectors. It converts the angular velocity into the angular momentum. For fixed rotation it is enough to consider this as a simple scalar multiplication. But for the complications associated with free rotation we need the full power of tensors to describe the dynamics.

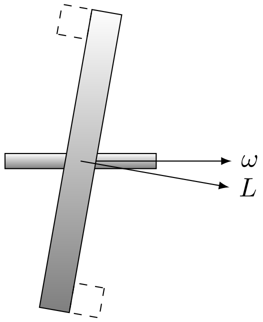


Figure 10.7: Example of dynamic balancing

Consider a wheel that is bent on its axle like Figure 10.7. The angular momentum L is aligned with the wheel, but the angular velocity ω is aligned with the axle. The axle must torque the angular momentum vector around the angular velocity. Newton's third law says that the wheel will also torque the axle. These extra torques cause the jarring vibrations experienced with such a bent axle.

One solution called **dynamic balancing** involves adding additional mass (indicated by the dashed squares) in order redistribute the mass and pull the angular momentum vector in line with the rotation. Mathematically, this balancing is done by examining the tensor components of the moment of inertia and figuring out where and how much mass is required to create stability.

This is free rotation without any external forces. If an external force is present it can cause **nutation** which is a slight vibration in the precession of the rotating object. This nutation is sometimes observable in a gyroscope as a slight up and down "nodding" as it precesses. In this case the nutation is caused by the weight of the gyroscope pulling down.

The rotation of the earth exhibits both precession and nutation. The earth is not quite spherical (bulges out at the equator) which causes a precession of its rotation—this means that the north pole is slowing moving away from the north star. The precession period is 26,000 years and was large enough to be noticed by the ancient astronomers. The nutation in the earth's rotation is much smaller in magnitude. The largest contribution is from the moon's gravitation and has a period of about 18.6 years.

Next week we will discuss celestial mechanics. We will need all of the mechanical tools we have developed so far to truly analyze the motion of the planets. We will talk about how to analyze the elliptical pattern of orbits and how this relates to energy and angular momentum. We will also see how these parameters are manipulated by space probes to move about the solar system. To top it off we will touch on how general relativity corrects Newton's law of gravity for extremely dense objects like black holes.

Lecture 11

Celestial Mechanics

Review lectures 6, 7, 9, and 10

We've already shown how Newton's law of gravity, equation (6.3) can be used to establish Kepler's third law (see page 35). We can now show that the second law follows from the conservation of angular momentum.

The moment of inertia for a particle rotating about an axis is simply mr^2 where r is the distance from the axis. The angular momentum of this particle is therefore

$$L = (mr^2)(\omega) = mrv_t \tag{11.1}$$

where we have used the fact that $v_t = r\omega$.

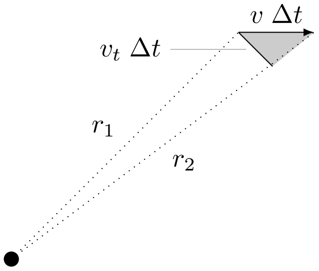


Figure 11.1: Kepler's second law and angular momentum

If we look at the position of the particle between two points in time we will get a diagram like Figure 11.1. The triangle defined by these positions has a height equal to $v_t\ \Delta t$ and we can say it has a base equal to r. So this area is given by

$$A = \tfrac{1}{2} r v_t\ \Delta t = \tfrac{1}{2}(L/m)\ \Delta t$$

Since the force of gravity is central, its line of action runs through the central point. This means it has no lever arm and produces no torque. As a consequence, the angular momentum is conserved. Therefore the area swept out by the radius vector is equal for equal time periods—which is Kepler's second law.

Kepler's first law is more difficult to derive. We'll skip the proof and move on to learning a few things about calculating with ellipses.

For problems in celestial mechanics, it is easiest to use polar coordinates. In order to consider problems relevant to Kepler's first law, we will consider an ellipse with its focus at the origin of a system of polar coordinates. In addition, assume the $\theta = 0°$ line is aligned with major axis of the ellipse like in Figure 11.2.

The point of closest approach for the planet is called **perihelion**. The suffix "helion" refers to the sun. If we are talking about orbiting the earth, the correct term would be perigee (the general term is periapsis). We will give this the label r_0.

On the opposite side of the ellipse we have the **aphelion** which is the point of farthest excursion. For the earth we say "apogee" and in general we say "apoasis". We will label this point r_1.

Both of these points lie on the major axis of the ellipse. So these three parameters are related by the following formula:

$$r_0 + r_1 = 2a \tag{11.2}$$

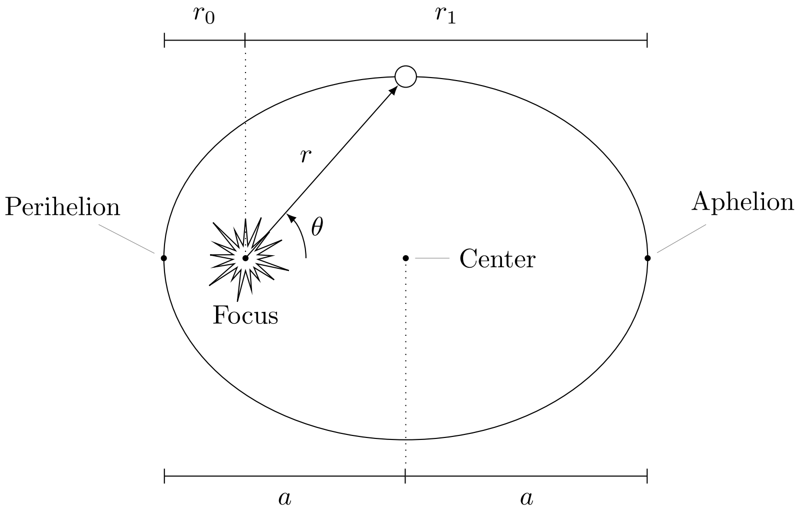


Figure 11.2: Key parameters and definitions for an ellipse

where a is the semi-major axis (i.e., half of the major axis).

The equation that describes the shape of the ellipse in polar coordinates is

$$r = r_0 \frac{1+e}{1+e\cos\theta} \tag{11.3}$$

where e is called the **eccentricity** of the ellipse. An eccentricity of zero corresponds to a perfect circle. Ellipses are characterized by an eccentricity less than one—the larger the number, the flatter the shape.[1] Notice that equation (11.3) does not involve time. This is merely an equation to describe the geometry of the path.

The more eccentric the ellipse, the farther the focus is from the center. This distance is half of $r_1 - r_0$. It happens that the formula for the eccentricity is

$$e = \frac{r_1 - r_0}{r_1 + r_0} \tag{11.4}$$

which is just this distance normalized by the semi-major axis of the ellipse.

The ellipse is characterized geometrically by the parameters a (the size of the ellipse) and e (the shape of the ellipse). Equations (11.2) and (11.4) show how to move back and forth between these parameters and the observable parameters r_0 and r_1. These facts are the bare minimum we need in order to understand the basics of celestial motion.

Now we are ready to discuss some dynamics. The potential energy function for gravity is given by equation (7.4). It is possible to reduce the two-dimensional motion of the planet to one dimension by using a rotating reference frame tied to the planet. Use of this frame will introduce a centrifugal force given by equation (9.6). We can use the conservation of angular momentum to eliminate the angular velocity from this expression. Since $L = mr\omega^2$, we have

$$F_{\text{cfg}} = \frac{L^2}{mr^3}$$

This is similar to the formula for gravity (6.3) and there is a corresponding "potential energy" associated with it. If we combine this with the potential energy for gravity, we have the following:

$$U = -\frac{GMm}{r} + \frac{L^2}{2mr^2} \tag{11.5}$$

This is called the **effective potential** and acts as the potential energy for the radial motion of the planet. (See Figure 11.3). Any orbital energy above the

[1] We can incorporate unbound orbits by allowing the eccentricity to be greater than one.

effective potential is kinetic. Remember this is the kinetic energy associated with the radial motion not the angular motion. The centrifugal term is the manifestation of the angular motion. In fact, using equation (11.1) one can see that the centrifugal "potential" can also be written at $\frac{1}{2}mv_t^2$. So this term can be seen as either potential or kinetic.

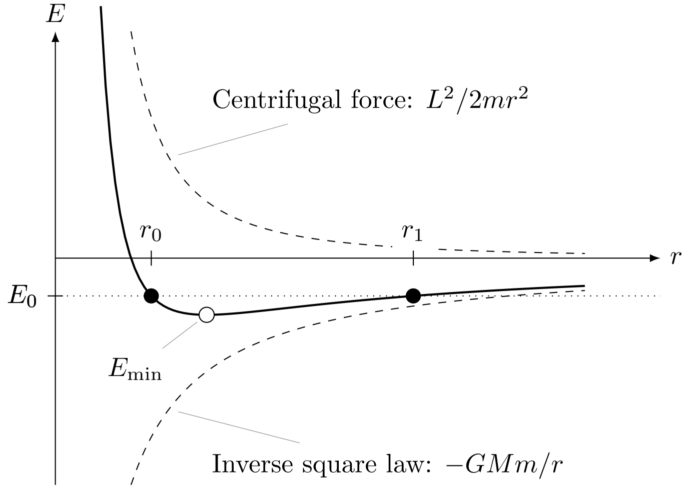


Figure 11.3: The effective potential for gravity

Note that the effective potential depends implicitly on the angular motion of the planet through the angular momentum L. The faster it rotates, the larger the centrifugal term, and the farther away will be the point of closest approach. The minimum point (E_{min}) on the effective potential corresponds to a purely circular orbit because the radius does not change—it has no kinetic energy of motion in the radial direction.

The effective potential gives us enough information to connect the motion of the planet to its dynamics. Specifically we can determine the relationship between the orbital energy and angular momentum (the dynamical parameters) to its perihelion and aphelion because they are the turning points on the energy diagram. We have already seen that these points are directly related to the eccentricity and semi-major axis of the ellipse (the orbital parameters).

The turning points are determined by setting the effective potential equal to the total orbital energy of the planet and solving for r. But the effective potential (11.5) makes it look like the orbit of the planet depends on its mass, which it does not. The acceleration due to gravity is independent of the mass of the object, so should the orbit of the planets. In order to emphasize this, let's introduce the "reduced" quantities $E' = E/m$ and $L' = L/m$. In fact, I'll even drop the primes and simply use the reduced versions because we rarely need to know the original ones. Multiplying both sides by $2mr^2$ and rearranging yields the quadratic equation we now need to solve:

$$2Er^2 + 2GMr - L^2 = 0$$

Determining both the eccentricity and the semi-major axis involve the sum and difference of the solutions to this equation.[2] For the sum we have:

$$r_1 + r_0 = -GM/E$$

and the difference is:

$$r_1 - r_0 = \sqrt{\left(\frac{GM}{2E}\right)^2 + \frac{L^2}{2E}}$$

The semi-major axis of the ellipse is half the sum of the perihelion and aphelion (11.2), so

$$a = -GM/2E \tag{11.6}$$

[2] For a general quadratic equation, $ax^2 + bx + c = 0$, the sum of the solutions is $-b/a$ and their difference is $\sqrt{(b/2a)^2 - (c/a)}$.

and the eccentricity is the ratio of the difference to the sum (11.4), so

$$e = \sqrt{\frac{1}{4} + \frac{EL^2}{2(GM)^2}} \tag{11.7}$$

These results are enough for us to start working some problems.

Consider a rocket satellite orbiting around the earth in a circular orbit with radius of 4.5×10^7 m. A sudden blast of the rocket motor increases the speed by 15% in the direction of motion. See Figure 11.4. Find (a) the maximum distance of the rocket, and (b) the eccentricity of the new orbit.

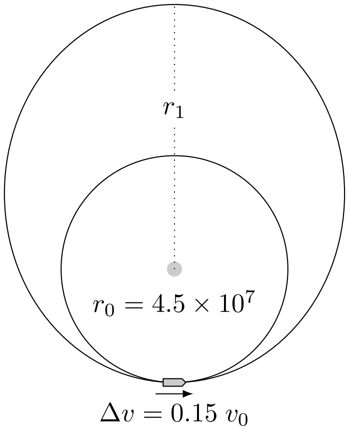


Figure 11.4: Rocket blast problem

Ever thought you'd be a rocket scientist? The quantity GM occurs frequently, so let's make a note of the product:

$$\begin{aligned} GM &= (6.673 \times 10^{-11})(5.9742 \times 10^{24}) \\ &= 3.9866 \times 10^{14} \end{aligned}$$

In this case M is the mass of the earth. Back in Lecture 6, we discussed Kepler's third law. In the process we derived the formula $GM = rv^2$ for circular orbits (see page 35). We can use this to determine the velocity of the satellite before the blast.

$$\begin{aligned} (3.9866 \times 10^{14}) &= (4.5 \times 10^7)(v_0)^2 \\ \implies v_0 &= 2976.4 \end{aligned}$$

The blast occurs in the direction of motion, so all of the Δv is tangential and the velocity after the blast must be

$$v = (1.15)(v_0) = 3422.9$$

Since the new orbit is not circular, the speed of the satellite will not always have this value. However, this point does now act as perigee for the new orbit (see Figure 11.4). Since we know the distance and velocity at this point, we can calculate the reduced energy and angular momentum.

There are a couple of ways to get at the total (reduced) energy, but perhaps the simplest is to simply take the sum of the (reduced) kinetic and potential energies:

$$\begin{aligned} E &= \tfrac{1}{2}v^2 - GM/r \\ &= \tfrac{1}{2}(3422.9)^2 - \frac{3.9866 \times 10^{14}}{4.5 \times 10^7} \\ &= -3.0010 \times 10^6 \end{aligned}$$

The energy is negative because the satellite is still in a bound orbit. The angular momentum is also straight-forward to calculate because at perigee (and apogee) all the velocity is tangential.

$$\begin{aligned} L &= rv_t \\ &= (4.5 \times 10^7)(3422.9) \\ &= 1.5403 \times 10^{11} \end{aligned}$$

Having calculated the dynamical parameters for the orbit, it remains to determine the orbital parameters. Using equation (11.6), we get

$$\begin{aligned} a &= -\frac{(3.9866 \times 10^{14})}{(2)(-3.0010 \times 10^6)} \\ &= 6.6421 \times 10^7 \end{aligned}$$

Now equation (11.2) yields the apogee:

$$r_1 = 8.7842 \times 10^7$$

And equation (11.4) yields the eccentricity:

$$e = 0.32251$$

So, the final answer is 8.8×10^7 meters for farthest excursion and the orbit has an eccentricity of 0.32.

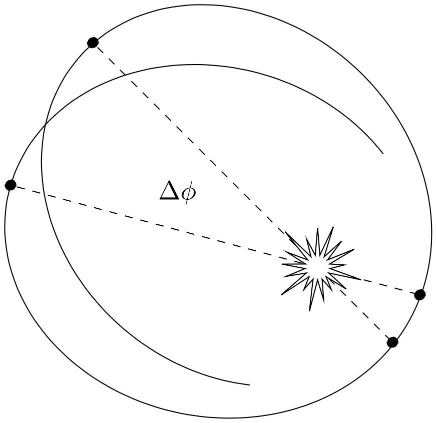


Figure 11.5: Measuring orbital precession

When we talk about the motion of the planets, Kepler's laws are a good first approximation—they describe the motion of each planet around the sun. What they don't take into account is the interaction between the planets. For example, the pull of Jupiter will cause each other planet to deviate slightly from a pure Keplerian ellipse.

In general, any slight deviation from the central inverse square law will manifest itself as a slight **orbital precession**. That is, the axis of the ellipse will itself slowly rotate causing the orbit to not quite close, like a spirograph. In Figure 11.5 this angle is denoted by measuring the movement of aphelion, although it is typically easier to measure the precession of perihelion.

With the improvement in observations and calculations over time, the orbits of all the planets were explained—including the discovery of Neptune and Pluto. By the middle of the 19th century, the slight irregularities and precession in Kepler's orbits were fully explained by Newton's law of gravity, except one. After taking into account all effects, there still remained a 43 arc-second per century precession in Mercury's orbit unaccounted for (this is a little over 7% of the total precession observed). This discrepancy was unresolved until 1916 and Einstein's general relativity.

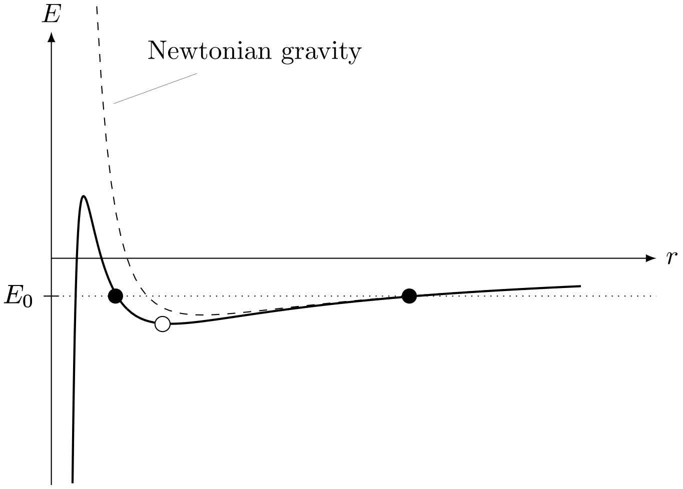


Figure 11.6: The effective potential in general relativity

Although general relativity requires a significant amount of math to fully appreciate, we can learn a bit about its consequences. For us, the easiest impact to see is on the effective potential. General relativity adds a new term to the calculation. The complete (reduced) effective potential is[3]

$$U = -\frac{GM}{r} + \frac{L^2}{2r^2} - \frac{GM}{c^2}\frac{L^2}{r^3} \tag{11.8}$$

This revised effective potential is plotted in Figure 11.6. The third term causes a precession which Einstein calculated as

$$\Delta\phi = \frac{6\pi}{a(1-e^2)}\frac{GM}{c^2} \tag{11.9}$$

where a and e are the semi-major axis and eccentricity of the orbit, respectively. This formula explains the anomalous precession of Mercury and was one of the three initial successes of Einstein's new theory of gravity.[4] The quantity GM/c^2

[3] This equation is exact—no slow speed or small mass approximation is necessary.

[4] The other two were gravitational redshift and the deflection of starlight.

is characteristic of the effects of general relativity.

You can see that the new term predicts a slightly stronger law of gravity than Newton's. Ultimately this comes from the fact that $E = mc^2$, so the gravitational field itself acts as a source of gravity. Both the equilibrium point and perihelion are pulled a little closer. But more dramatically, if the incoming object has sufficient energy the gravitation well will capture it regardless of the centrifugal force.

This point of no return is called the Schwarzschild radius, $r_s = 2GM/c^2$. If an object is crushed below its own Schwarzschild radius, it will not be able to prevent its own collapse—a **black hole** will result. That is why this distance is traditionally called the radius of a black hole even though it really does not have any extension. Notice that black holes do not require a lot of mass—what they require is an extremely dense mass.

Of course, black holes are only one of the dramatic predictions of general relativity. We don't have time to cover them all: curvature of space-time, gravitational waves, expansion of the universe, the existence of "dark energy".[5] General relativity has withstood the test of time and stands today as one of the two pillars of fundamental physics.[6] It also represents the end the road for this term regarding the motion of simple objects.

[5] See Lecture 13 and 26 for more info.

[6] The standard model is the other pillar which we will cover in Lecture 30.

Lecture 12

Harmonic Motion

Read sections 10.1–10.6

In Lecture 7 we discussed the fact that nearly all physical systems sit in a state of stable equilibrium (see page 41). A state is in stable equilibrium if the internal forces of the system tend to push it back into that state. This must be true along all the degrees of freedom for the system. For example, a rigid object is in equilibrium only if both the forces and the torques in the system are balanced (see Lecture 10).

If the system is somehow displaced from this point of equilibrium, it will be pushed back. As the system returns to this equilibrium state, its inertia will tend to carry it through and past equilibrium. So the system touches its equilibrium only to move away in the opposite direction. Eventually the forces in the system become large enough to overcome the inertia of the motion, bring it to a stop, and push it back toward equilibrium. Eventually the system will be back to the displaced state in which it started. And the cycle repeats over and over. This back-and-forth motion is called **vibration** and a system can vibrate along any of its degrees of freedom, whether that is a spring, a string, the surface of a drum, a metal beam, a column of air, or the electromagnetic field.

The maximum extent of the displacement is called the **amplitude** of the motion and the time it takes for one complete cycle is called the **period** of the motion. It is no coincidence that this is the same term we used in Lecture 6. The reciprocal of the period is called **frequency** and its unit is the **hertz**—so when the vibration of a system is said to have a frequency of 10 hertz, we mean that its motion repeats 10 times each second.

The simplest example of a vibrating system is a spring with one end attached to a fixed point and the other end attached to a certain mass. We will assume the spring is oriented horizontally to ignore gravity and we will assume the mass slides without friction (see Figure 12.1).

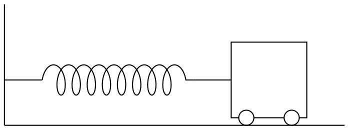


Figure 12.1: The simplest vibrating system: an ideal spring

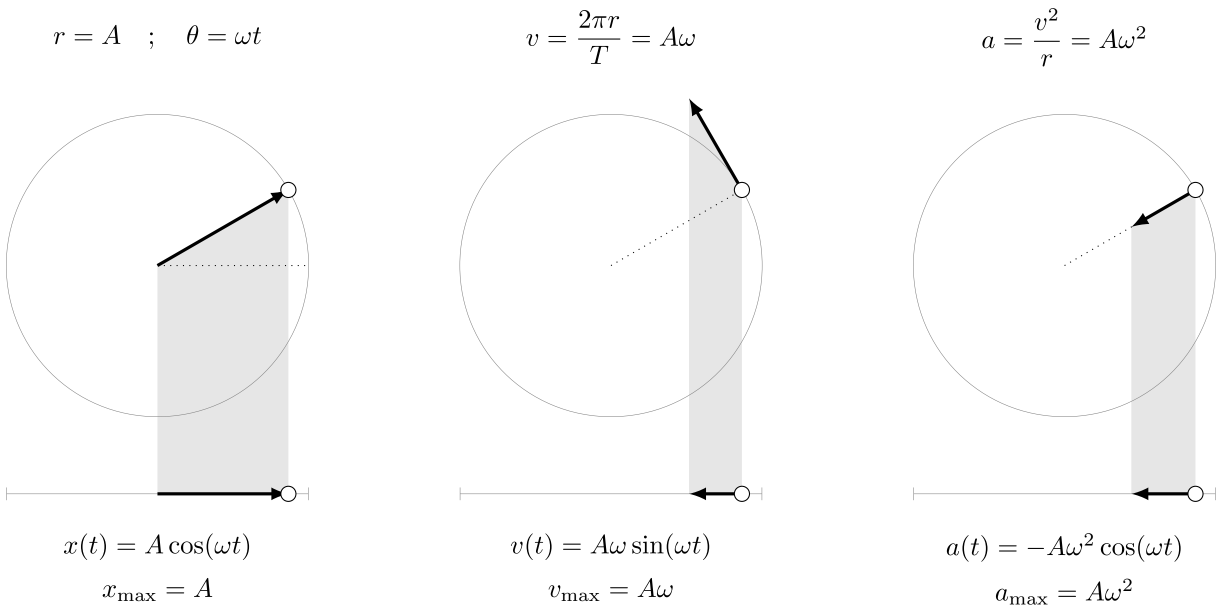
This system is in equilibrium when the spring is not stretched. We define the position of the mass to be zero at this point. Any displacement to the right will be positive, and any displacement to the left will be negative. In either case, the spring will exert a force that will tend to restore the mass to equilibrium. In other words, the mass will experience a force that is in opposition to its displacement. The spring is considered an **ideal spring** if the magnitude of this force is exactly proportional to the displacement. In symbols, the restorative force of an ideal spring is

$$F = -kx \tag{12.1}$$

where k is called the **spring constant** and is measured in newtons-per-meter and this equation is frequently called **Hooke's Law**. The motion of a mass subject to this force law is called **simple harmonic motion**.

When our simple mass-and-spring system is vibrating, the restorative force will change in time as it responds to the motion of the mass. This means that the constant acceleration equations from Lecture 3 will not be applicable to this system. We need a different way to analyze this kind of vibrating motion. This is similar to the situation we faced in Lecture 6 with uniform circular motion, except we now know the force but not the mathematics of the motion.

All is not lost. By a fortunate fluke, we can leverage our learnings from Lecture 6 to help us now. Consider the projection of uniform circular motion onto the x-axis. Figure 12.2 shows the essential correspondences.



$r = A \quad ; \quad \theta = \omega t$
$v = \frac{2\pi r}{T} = A\omega$
$a = \frac{v^2}{r} = A\omega^2$
$x(t) = A\cos(\omega t)$
$x_{\text{max}} = A$
$v(t) = A\omega\sin(\omega t)$
$v_{\text{max}} = A\omega$
$a(t) = -A\omega^2\cos(\omega t)$
$a_{\text{max}} = A\omega^2$

Figure 12.2: Uniform circular motion versus simple harmonic motion

The position of the particle in uniform circular motion is specified by its radius and its angle. Since the motion is uniform, the angle increases at a constant rate, so we have $\theta = \omega t$, where ω is the angular speed of the particle. The projection of this position onto the x-axis involves a right triangle. The diagrams on the bottom of Figure 12.2 show this projection: the position of this particle is given by

$$x(t) = A\cos\omega t \tag{12.2}$$

where we have also relabeled the radius as A since this is also the amplitude of the projected motion.

In Lecture 6 we found that the acceleration in uniform circular motion is directed in toward the center. Since this is directly opposed to the position vector, the same is true in the projection. That is, the acceleration is opposed and proportional to the position. This is precisely the kind of connection we expect with an ideal spring.

In fact, we can say more. We know that the speed of the uniform circular motion is related to the radius and period of the motion via $v = 2\pi r/T$. By definition, the angular speed is the 2π radians divided by the time period T, so we can rewrite this as $v = A\omega$. (This should also remind you of the formula for tangential velocity (9.3) in Lecture 9.) Since this velocity vector is pointing perpendicular to the position vector, the projection of the velocity is given by

$$v(t) = A\omega\sin\omega t \tag{12.3}$$

The formula for the acceleration in uniform circular motion is from equation (6.1). Using our work so far, we can rewrite this as $a = A\omega^2$. Since the projection is opposed to the displacement, the formula for the acceleration of simple harmonic motion is given by

$$a(t) = -A\omega^2 \cos \omega t \tag{12.4}$$

We can get to this same spot without the analogy to circular motion using a bit of calculus. Those of you familiar with those techniques will recognize the connections between the sinusoidal functions here. In either case, equations (12.2)–(12.4) describe the kinematics of simple harmonic motion.

In the preceding, we have assumed the particle begins at its point of maximum extension. We can remove this restriction by adding an extra term to equation (12.2):

$$x(t) = A\cos(\omega t + \phi) \tag{12.5}$$

The quantity ϕ is called the **phase shift** and will become important when we discuss wave interference in Lecture 19. Any system which obeys equation (12.5) is called a **simple harmonic oscillator**.

It remains to connect these considerations with the parameters of our ideal spring system. We can put together Hooke's law (12.1) with Newton's second law (5.1) to get

$$-kx = ma$$

After plugging in the results from equations (12.2) and (12.4) and simplifying we get

$$k = m\omega^2 \tag{12.6}$$

This little equation is an important summary of the dynamics of the ideal spring. In general, any vibrating system will have a similar equation connecting the parameters of the system to the frequency of its motion. The value of ω is called **angular frequency** to distinguish it from the vibration frequency defined earlier. Because there are 2π radians in each cycle, the two quantities are simply connected by the equation $\omega = 2\pi f$. So, we can write the vibration frequency of the simple harmonic motion in terms of the spring constant and the mass:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{12.7}$$

This is called the **natural frequency** of the ideal spring. For any system in stable equilibrium, each degree of freedom will have its own natural frequency.

The amplitude in equation (12.5) is related to the energy in the system. Hooke's law (12.1) is conservative and the potential energy associated with it is[1]

$$PE_{\text{spr}} = \tfrac{1}{2}kx^2 \tag{12.8}$$

In general, the total energy of the ideal spring system is sum of the kinetic energy of the mass and the potential energy in the spring. But when the system is at maximum extension the velocity of the mass is zero (for a moment). Since $x = A$ at this point, we can write

$$E_{\text{tot}} = \tfrac{1}{2}kA^2 \tag{12.9}$$

The fact that the energy of vibration is proportional to the square of the amplitude of the vibration explains why laser light is so much more powerful than regular light (see Lecture 27).

Hooke's law (12.1) is an idealization for a real spring, and even more so for complex physical systems. However, the forces maintaining any state of stable equilibrium are approximately linear for small displacements. One easy way to see this is by investigating the energy diagram. For a system to be in stable equilibrium, the potential energy curve must create a trough or potential well. The point of equilibrium is at the bottom of this well. If one were to fix the vertex

[1] This can be seen from the definition of work (7.1)

$$\Delta W = F\Delta x$$

In this case we have to be careful because the force is not constant. But the work does increase at a constant rate, so we use the trick from Lecture 3 where the average value is equal to the average of the ends. In this case, the work done at the beginning is zero because there is no net force at equilibrium. As we pull on the spring, the work required increases. At full extension, the work required is $W = (-kx)(x)$. Taking the average introduces a factor of one-half.

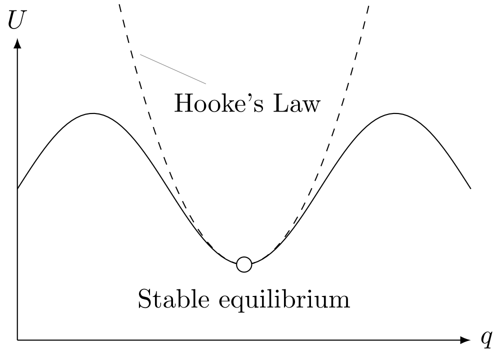


Figure 12.3: A parabola can be made to fit into any potential energy well

of a parabola to this point and fit its curvature to the curvature of the potential energy function, one might get a diagram like Figure 12.3.

This shows that any potential well can be approximated by a parabola associated with some version of Hooke's law (12.1).[2] The project for this term uses this trick to calculate an estimate for the anomalous precession of Mercury due to general relativity.

An important example of a non-linear system is the pendulum. In this case we consider our displacement to be the angle of the pendulum from vertical. The weight of the pendulum bob creates a torque that has a tendency to push the pendulum back to center (see Figure 12.4).

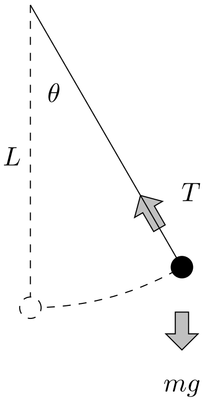


Figure 12.4: Forces on a pendulum

The tension from the string has no lever arm, so it introduces no torque into the system. The tangential component of the weight is simply $-mg\sin\theta$, so the total torque on the pendulum is

$$\tau = -mgL\sin\theta$$

Newton's second law for rotation states that this will cause an angular acceleration:

$$-mgL\sin\theta = mL^2\alpha$$

where I have used the fact that the moment of inertia for this simple pendulum is mL^2. Now if the angle involved is small, we can use the approximation that $\sin\theta \approx \theta$ (when θ is measured in radians).[3] We can also absorb an L on both sides to convert the angles into tangential variables. We get

$$-gs = La$$

where s represents the arc-length of the pendulum's swing. This equation is similar to $-kx = ma$ for an ideal spring. So, we can immediately write the frequency for the pendulum swing as

$$f = \frac{1}{2\pi}\sqrt{\frac{g}{L}} \tag{12.10}$$

Remember, this equation is only valid when the angular displacement is small.

The ideal spring is a good model for any vibration. We can improve the model by adding an element of friction to the system. There are different ways to do this, but one very common way is by introducing a drag term which is proportional to the velocity of the system:

$$F_{\text{drag}} = -cv \tag{12.11}$$

where c is the "stiffness", or viscosity of the drag. This is the type of friction an object experiences as it pushes slowly through a fluid. The shock absorbers in your car use this kind of a set up to minimize vibration.

This damping effect introduces a non-conservative force into the system which will reduce the amplitude of the motion over time. Since the amplitude is related

[2] The curvature of the parabola is related to the second derivative of the potential energy function.

[3] Small means less than 10°, or less than 0.1 radian. This is the point in the derivation where we replace the natural potential energy function with the ideal parabola for simple harmonic motion.

to the energy of this system, this reflects the fact that energy is being lost through friction. The equation for the amplitude is

$$A(t) = A_0 \exp(-\gamma t)$$

where $\gamma = c/2m$. The displacement of the damped oscillator over time is graphed in Figure 12.5.

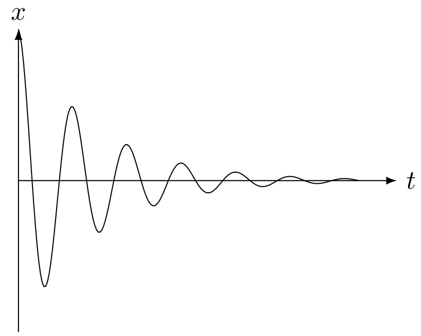


Figure 12.5: The motion of a damped harmonic oscillator

When $\gamma = \omega$, the system is said to be **critically damped**. This is the circumstance which brings the system back to equilibrium the fastest. If the system is under-damped ($\gamma < \omega$), then it will oscillate a bit before being brought to rest, as described above. If the system is over-damped ($\gamma > \omega$), then the viscosity will be so thick it will actually drag against the system being brought to rest.

Finally, we have one last topic to consider: forced vibrations. If a damped system is driven by an external vibration, the motion of the system will eventually match the frequency of the driver. Initially it will move through a temporary ($\Delta t > 1/\gamma$) **transient state** until it reaches a **steady state** of sinusoidal motion. This steady state will have an amplitude which will depend upon both the frequency of the external driver and the natural frequency of the system.

What happens is that the external vibration does work on the system when the force is aligned with the displacement of the system. When the frequency of the external vibration is matched with the natural frequency of the system, the work done in each cycle accumulates. It does not take long for system to absorb a lot of energy from the external source even if the magnitude of the source is small. Eventually the system is moving fast enough that the damping force (proportional to this speed) increases to drain the energy that is being absorbed. This steady state is called **resonance**.

The typical power absorption for a driven harmonic oscillator is plotted in Figure 12.6. The peak power absorption occurs at the natural frequency of the system. The width of the peak is related to the “stiffness” of the damping: at one-half of the maximum power absorption it is equal to 2γ.

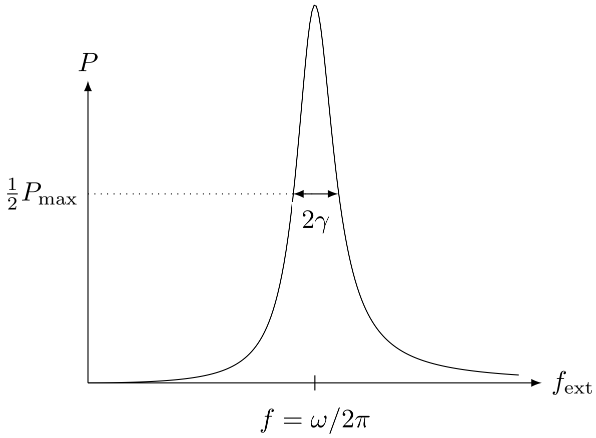


Figure 12.6: Power absorption in a driven harmonic oscillator

Resonance can be used to explain the scattering of light (why the sky is blue), how microwaves work, and is important in certain electrical circuits. The quality factor of an electrical circuit is defined as the ratio of the resonant frequency and the half-maximum width of the power absorption curve. A sharp peak will make a good radio receiver as it will only react to a particular range of radio frequencies. A quality factor between 10 and 100 is common. For our ideal spring, this quality factor would be

$$Q = \frac{\omega}{2\gamma} = \sqrt{\frac{km}{c^2}}$$

which shows how this one metric incorporates all the essential parameters of the system.[4]

[4] The quality factor for a resonating circuit is $\sqrt{L/CR^2}$—see Lecture 25.

Next lecture we will extend this notion of things that stretch to cover the topic of the elasticity of solids. We will find that forces create stress in an elastic solid and that this stress can be decomposed into three essential components. Stress causes strain, and Hooke's law will reappear in the relationship between them. We will take a quick overview of the various ways real objects deform and react to various stresses.

The idea of stress also offers us a segue to discuss general relativity. We will see that the essential effect of gravity is to create stress in the space-time continuum: which is witnessed by the presence of the ocean tides. We will finish by discussing whether it is possible to avoid being crushed while falling into a black hole.

Lecture 13

Elasticity

Read sections 10.7–10.8

As discussed in Lecture 1, physics is often described as the study of matter and energy. So far, we have focused on the dynamics of motion. Now we turn to a study of the properties of matter—this will occupy us for the next few lectures. The most basic classification of material things is by their phase: whether they are solid, liquid, or gaseous. At the microscopic level, these phases are distinguished by the strength of the inter-molecular interaction. The solid phase is such that the configuration of the molecules are essentially locked together. They may vibrate in place, but there is no bulk motion relative to one another.

In the fluid phases, the molecules can flow around one another and move freely. Both liquids and gases are considered fluids because of their ability to flow. The two are distinguished by the fact that the molecules in liquids stay relatively close together (they constantly "touch"), but in gases they fly about nearly free of interaction.

In this lecture we will discuss solids and their elastic properties. Frequently a solid is depicted as a bunch of molecules connected to one another by springs (see Figure 13.1). The springs are meant to represent the inter-atomic forces between the molecules. As we discovered in the previous lecture, the forces in any system near stable equilibrium can be approximated by ideal springs, so this picture is not as unrealistic as it may first appear.

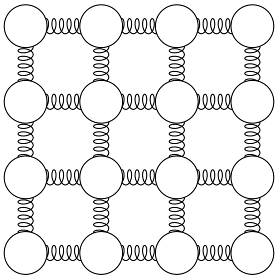


Figure 13.1: A solid as a collection of molecules connected by springs

This conceptual model also shows the main property of a solid: its elasticity. When any force is applied to a solid, it will respond through deformation and will resist the force applied. The details of this deformation can be extremely complicated—as can be seen just be considering the number of degrees of freedom available to all the molecules.

The simplest approach is to consider a solid rod. This is somewhat like a one-dimensional system. Assume the rod is fixed at one end and a force is applied to the other. In general, the force will have two components: along the length of the bar and across the length of the bar. The first is called **normal stress**[1] and the second **shear stress**. The reason we use the word **stress** rather than force here is that we expect all the forces to be in balance—we are interested not in the net force which produces motion, but the stable combination of forces which produces deformation. Stress is what's left after the net force has been calculated.

Normal stress will either stretch (also known as tension) or compress the solid (see Figure 13.2).[2] The resulting deformation is called the **strain** in the object and is defined as the amount of deformation divided by the original length of the bar. The strain is a unit-less number: the percent change in the original length. If the force is distributed over a large area, the resulting strain will be less. So we are

[1] "Normal" because the force is perpendicular to the cross-section of the bar.

[2] In three dimensions we can distinguish between tension and compression by the fact that tension will cause expansion in one dimension and contraction in the other two. Compressive forces will cause expansion in two dimensions and contraction in one.

really talking about the pressure from the normal force causing the deformation. When the displacement involved is small, the controlling equation is

$$\frac{F}{A} = Y\left(\frac{\Delta L}{L_0}\right) \tag{13.1}$$

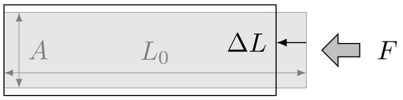


Figure 13.2: A solid rod under normal stress

Notice how this is similar in form to Hooke's law. The quantity Y is called **Young's modulus** and is on the order of 10^{11} for metals in SI units (indicating enormous forces are required to create a significant deformation).

Shear stress is similar, except now we are interested in quantifying the deflection of the bar (see Figure 13.3). The formula for shear stress is nearly the same:

$$\frac{F}{A} = S\left(\frac{\Delta x}{L_0}\right) \tag{13.2}$$

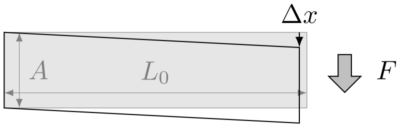


Figure 13.3: A solid rod under shear stress

except now the strain is quantified as the ratio of the deflection to the bar's length. The proportionality constant S is called the **shear modulus**.

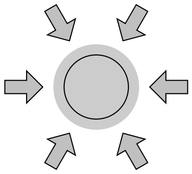


Figure 13.4: An object under compressive pressure

We have not exhausted all the ways a solid can be put under stress. You may notice that in Figure 13.2 there is a slight expansion along the cross-sectional area. The effect is exaggerated in the diagram, but the point is that normal stress will not change the volume of the object. When one part is squeezed, another will bulge. However, it is possible for forces to completely surround the object like in Figure 13.4. We will call this **pressure stress**.

In a way you can consider this a special case of normal stress, but we only take the average pressure in all directions. In other words, we break the normal stress into two components: (1) the average value from all directions (the pressure stress) which changes the volume of the object and (2) the remainder (the normal stress proper) which merely distorts the shape of the object.

The formula for pressure stress is

$$\Delta P = -B\left(\frac{\Delta V}{V_0}\right) \tag{13.3}$$

where B is called the **bulk modulus** of the material. The negative sign is here because an increase in pressure will cause a decrease in volume.

In the formulas (13.1), (13.2), and (13.3), the forces involved must be small. In fact, they are all similar in form to Hooke's law (12.1) and are sometimes referred to as such. When the forces become larger, these equations cease to work and tend to over-predict the stress required for a particular level of strain. In other words, the object tends to break apart. There are three typical limits of strain for a given object:

- Proportionality limit
- Elastic limit
- Deformation limit

These are highlighted in Figure 13.5. Up to the proportionality limit, Hooke's law is valid and each increment of stress causes a consistent increment of strain. Past the proportionality limit and up to the elastic limit, the object is plastic, meaning that when the stress is removed the object still will conform to its original shape. If the stress exceeds the elastic limit, the deformation is permanent and if the stress goes beyond the deformation limit, the object will break. Clearly understanding these limits is important in the design and construction of all kinds of objects.

Now, in general, an object will be under the influence of a variety of forces in multiple directions at a variety of points. One way to deal with this complexity is

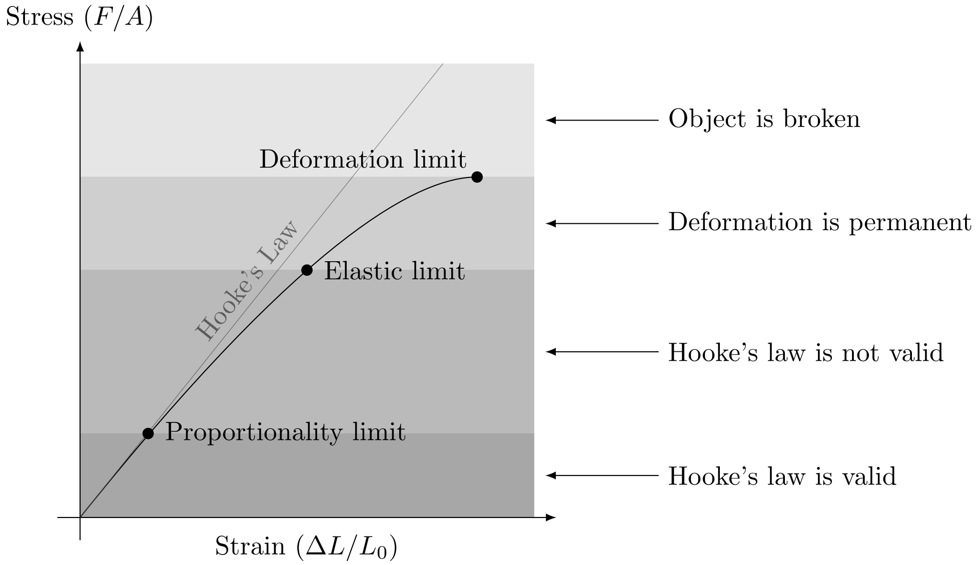


Figure 13.5: Typical stress-strain relationship for a solid object

to break the object into a (large) number of pieces. The simplest way to do this is by splitting it into little cubes. Each cube interacts with the others through the forces along their sides. We deliberately make the cubes so small that we are able to assume the forces are constant along each face. Also, since the cubes are small we can apply Hooke's law to each: in other words, the strain in each cube will be directly proportional to the forces on the sides. Over the volume of the object, the strain within the cubes accumulate and gives us the total deformation and deflection in the shape of the object. This approach is called **continuum mechanics**.

Consider the forces on one of these elemental cubes. Since we are only interested in the stress and not the bulk motion of the cube, we assume the forces are in equilibrium. This means that the forces on opposite sides of the cube are equal and opposite, so we only need consider one side of each pairs. For extreme simplicity, we will consider the two-dimensional "cube" shown in Figure 13.6. The Greek letter σ is traditionally used to denote stress. In our case $\vec{\sigma}(\hat{x})$ represents the force of stress along the face with the x-direction as its normal. I've written it this way to emphasize the fact that stress σ is a tensor: it maps each direction vector into the force of stress along that direction.[3]

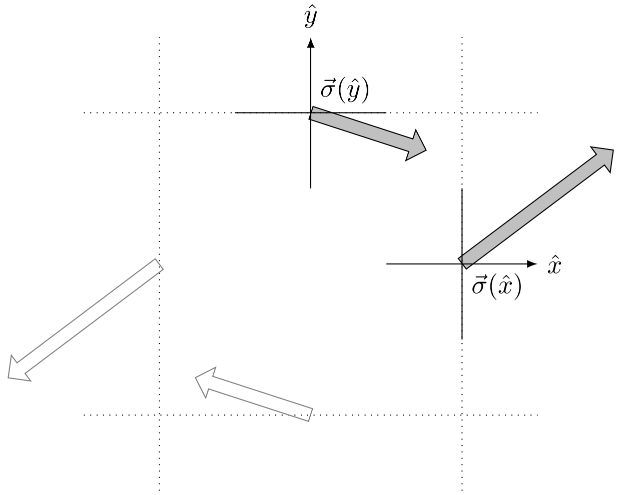


Figure 13.6: Forces of stress on a cubical element

We can go a bit further by breaking these stress vectors into their components.

[3] Just because it is a function between vectors is not enough for it to be a tensor. It must also be linear. One could show this by rotating the cube an comparing the results with the original orientation.

This is shown in Figure 13.7. Each component has a double subscript. The first letter denotes which cube face we are referring to and the second letter denotes the direction of the component. For example, σ_{xy} represents the y-component of the stress force $\vec{\sigma}(\hat{x})$. There are several advantages with this **index notation**. One of them is that we can easily write that the stress tensor is symmetric, i.e., $\sigma_{xy} = \sigma_{yx}$. This is because these are both the tangential components of their forces and produce torque. But the cube is in equilibrium, so these torques must cancel. This will only happen if these components are equal.

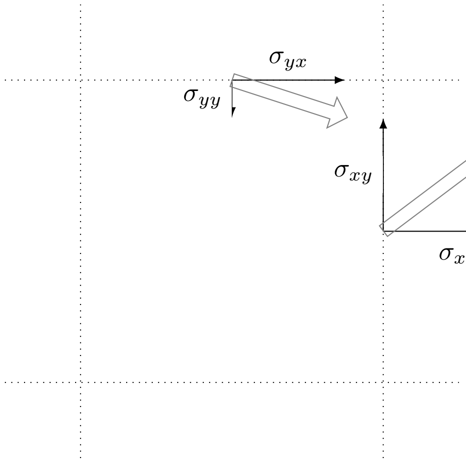


Figure 13.7: The components of the stress on a cubical element

We can take this even further by busting these components into pieces that emphasize their normal, shear, and pressure stress dynamics. We have already seen that the shear components are equal by definition, so that is done. The pressure piece is simply the average of the two normal components. These pressure pieces are also equal by definition. The remaining normal components are also equal (though opposite) because they are their difference from the average. This means that we are able to characterize the stress created by arbitrary forces using just three numbers related to the three components of stress we introduced at the beginning of this lecture. Figure 13.8 shows these "canonical" components for the example in Figure 13.6.

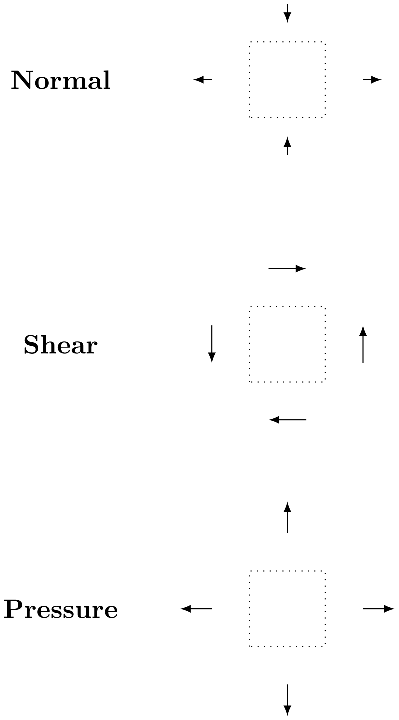


Figure 13.8: The three canonical components of stress for Figure 13.7

We can even talk about stress on an astronomic scale. The ocean tides result from the stress induced by the gravitational pull of the moon (and the sun). The gravitational force creates a normal stress on the oceans which tend to expand the water along the line between the centers of the earth and moon and compress the water perpendicular to this line.

The expansion can be understood because the inverse square law means that the force of gravity on the near side to the moon will be slightly larger than at the center, which will be slightly larger than on the opposite side. If we label the distance between the earth and moon d and the radius of the earth as r, we have

$$F_{\text{near}} = \frac{GMm}{(r-d)^2}$$

for the force of gravity on the near side. The force on the far side is similar with $r+d$ in the denominator instead.

We can rewrite this equation as

$$F_{\text{near}} = \left(\frac{GMm}{d^2}\right)(1 - r/d)^{-2}$$

The reason to do this is that we can use the binomial theorem (1.3) on the second factor because r/d is so very small. We have

$$(1 - r/d)^{-2} \approx 1 + 2r/d$$

The stress on the earth's oceans is given by the difference between this force and that at the center of the earth. For both the near and far sides this is the same in magnitude. We have:

$$\sigma_{xx} = \left(\frac{GMm}{d^2}\right)\frac{2r}{d}$$

where the x-direction is aligned along the earth-moon line.

We expect compression on in the y-direction, but the reasoning is different. In this case, we are looking for the component of the forces on the sides of the earth directed toward the center of the earth. The triangle involved in these components is similar to the triangle formed by the geometry of the earth-moon system (refer to Figure 13.9). This means we can write

$$\frac{F}{F_{\text{grav}}} = \frac{r}{d_{\text{side}}}$$

because in similar triangles the ratio of similar sides are equal.

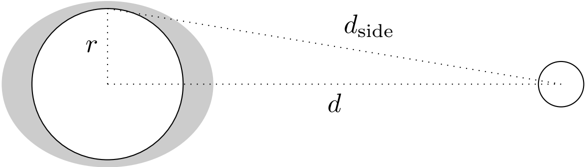


Figure 13.9: Tides on the earth caused by the gravitational force of the moon

The force of gravity F_{grav} on the side of the earth is proportional to the inverse square of the distance d_{side}, or

$$F_{\text{grav}} = \frac{GMm}{d^2 + r^2}$$

However, by using the binomial theorem earlier, we have implicitly agreed to completely ignore r^2 relative to d^2 since $r \ll d$.[4] By similar logic we can now write $d_{\text{side}} = d$.

This means that the normal stress across the earth-moon line is

$$\sigma_{yy} = -\left(\frac{GMm}{d^2}\right)\frac{r}{d}$$

where I have introduced the negative sign since the stress is compressive along this component.

The gravitational field introduces no shear stress, so this completely determines the tidal stress from the moon.

In classical mechanics we characterize the influence of gravity by Newton's inverse square law (6.3). But in Einstein's general theory of relativity we use these tidal stresses to define the influence of gravity. There are two reasons for this. The first is that we need a field theory—one that does not involve action-at-a-distance like Newton's law. The gravitational effect must propagate through the field at the speed of light rather than being instantaneous. The second reason is that there is always an inertial frame in which the force of gravity can be eliminated—this is the equivalence principle (see page 26). But even in a free-fall frame the tidal effects of gravity are present. In a way the tidal effects are more "real" than the force of gravity itself.

So in general relativity, gravity propagates through space as waves of stress in the fabric of space-time. In fact, gravity waves (if detected) will be measured through their extremely faint normal stresses in objects. Only waves from dramatic astronomic events like a supernova will be detectable directly, but there is already indirect evidence of the existence of gravity waves.

[4] For example, if $x = 10^{-6}$ is barely in our measurement range then $x^2 = 10^{-12}$ is certainly not.

As one falls into the gravitational well of an object, the free-fall frame experiences tidal forces similar to those from the moon. If the object is a black hole (crushed beyond is Schwarzschild radius—see Lecture 11), these tidal effects increase without limit. As one is pulled inexorably toward the central pit of destruction, the tidal forces will pull one into a long string of spaghetti.

This situation is unavoidable if the black hole is not rotating. If it does rotate (which almost all will since angular momentum is conserved), the **event horizon** that surrounds the black hole at its Schwarzschild radius splits in two. And it becomes possible to pass into the event horizon without being doomed to being crushed in the central singularity (the so-called **ergosphere**). In fact, there is even reason to believe we can extract energy for the rotation of the black hole by exploiting this ergosphere—this is called a Penrose process. There is even speculation that this ergosphere may offer an ability to travel through a black hole to another region of space-time—the **worm hole** made famous in many science fiction stories nowadays. But that discussion lies far afield from this lecture.

Next week we will move on from solids to discuss the dynamics of fluids. As mentioned earlier, these thoughts will apply to both liquids and gases. Initially we will discuss hydrostatics: the properties of fluids at rest. The main result will be a determination of the buoyant force in any fluid. After that we move to hydrodynamics, or fluids in motion. We will introduce some new concepts to speak quantitatively about fluid flow and characterize the different ways fluids move. Bernoulli's equation will be introduced as the main equation for hydrodynamics. We will touch on viscosity which will introduce some realism into our discussion. Finally we will touch on the topic of deterministic chaos which was first discovered in the study the weather by fluid modeling of the atmosphere.

Lecture 14

Fluids

Read sections 11.1–11.11

Of the three phases of matter, liquids and solids are distinguished by their ability to flow. Liquids flow downward under their own weight to take the shape of their container while gases flow by expanding to completely fill their container. The ability to transport mass from one point to another is the main thing that separates fluids from solids.

This distinction also shows up in how fluids react to the three types of stress introduced in the last lecture. Fluids don't break. In fact, fluids can't really support any normal or shear stress. Although a solid will deform and resist the stress, a fluid will simply react by moving out of the way. An ideal fluid will flow instantly, but any real fluid will have a bit of reaction time related to its viscosity. This means that pressure is the only kind of stress that a fluid at rest will support.

In fact, a fluid at rest won't even support a pressure differential. The fluid will move from high to low pressure. If the pressure increases anywhere in the fluid, this increase will distribute evenly throughout the fluid. This is called **Pascal's principle** and is what makes most hydraulics work. The hydraulic press is a simple example of this where a force on one end can be used to support a much larger force on the other (see Figure 14.1). Since the pressure must be equal and pressure is force divided by area, a larger area will deliver a larger force. The hydraulic press is a machine with an ideal mechanical advantage equal to the ratio of the area on either side of the press.

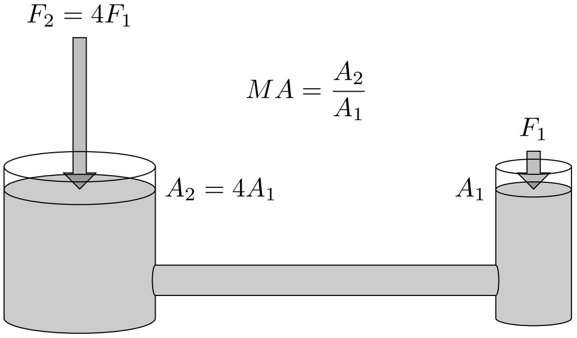


Figure 14.1: The mechanical advantage of the hydraulic press

The statement I made earlier when I said that a fluid will not support a pressure differential was not quite true. In the presence of gravity (or any long-range force for that matter), the fluid below must support the fluid above.[1] In fact, the pressure differential required is pretty easy to calculate. Imagine a rectangular box of height h with a horizontal cross-section of area A (see Figure 14.2).

[1] Obviously this requires a container of some sort. Fluid doesn't just pile up like sand.

We will assume the mass density[2] ρ of the fluid is constant, so the total mass of this imaginary box of fluid is

$$m = \rho V = \rho A h$$

The pressure on the bottom of the box must exceed the pressure at the top because it is holding up the weight of the fluid in the box. The total force from the fluid on the bottom will be $F_2 = P_2 A$. This force must balance both the pressure at the top of the box and also the weight of the box. Thus,

$$P_2 A = P_1 A + \rho A h g$$

When we cancel the area and write in terms of a pressure differential we get:

$$\Delta P = \rho g h \tag{14.1}$$

which I like to call the **hydrostatic equation**.

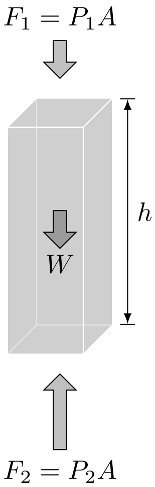


Figure 14.2: Deriving the hydrostatic equation

This reasoning provides us with a method for calculating the force of buoyancy in a fluid. This **buoyant force** is precisely this pressure differential multiplied by the cross sectional area. In other words

$$B = \rho g V \tag{14.2}$$

where B is the force of buoyancy, ρ is the density of the fluid, and V is the total volume displaced. This equation is sometimes called **Archimedes' principle**. If the density of the submerged object is greater than that of the fluid, its weight will be larger than the buoyant force. The net force will be down and the object will sink. If the density of the submerged object is less than the fluid, the object will rise until it breaks the surface of the fluid (presuming there is one). After that point, the displaced volume will get smaller until equilibrium is achieved between $\rho g V$ and the weight—the object floats.

Consider a swimming pool. What is the water pressure one meter below the surface? Well, according to equation (14.1), the answer is 9800 pascals (one **pascal** is equal to one newton per meter squared), or 9.8 kPa. But this is the pressure differential—what is the pressure at the top of the pool? Hint: It's not zero.

Yes, the pool is also supporting the weight of all the air above it. The atmosphere also acts like a fluid and this is the source of atmospheric pressure. We are all like fish swimming is a sea of nitrogen and oxygen gas. The typical value for atmospheric pressure near sea level is 101.3 kPa. So this is the pressure at the top of the swimming pool. The bottom of the swimming pool will have a absolute pressure of 111.1 kPa.

The pervasiveness of the air makes it easy to ignore. In fact, many tools designed to measure pressure (like a tire gauge) are calibrated to register zero at atmospheric pressure. These tools are said to measure **gauge pressure** which is simply the true absolute pressure minus one atmosphere's worth of pressure. Usually we work with pressure differentials so this issue washes out. On the other hand, occasionally we do need to know the absolute pressure so it's important to be aware of this distinction.

By the way, atmospheric pressure is how a straw works. It's often said that sucking on the straw creates a vacuum and that the vacuum sucks the liquid up the straw. This is not the right: the surrounding air pressure pushes the fluid up the straw. A straw won't work in the vacuum of space.

One last point about atmospheric pressure. Even if we ignore the fact that it is constantly in motion, this pressure does not obey equation (14.1). This is because in it we have assumed the density to be constant.[3] Which may be true for a liquid, but for a gas the density generally depends on its pressure (see the ideal gas law

[2] Remember that mass density is defined as the mass divided by the volume of the substance. Later we will see the density defined for other quantities like the number of particles or charge.

[3] We have also assumed the gravitational force mg to be constant.

(16.1) in Lecture 16). Taking this into account (with a bit of calculus), the correct equation[4] is

$$P = P_0 \exp(-kz) \tag{14.3}$$

where k is related to the molar mass and temperature of the gas.

For small heights, it is still appropriate to use equation (14.1) for a gas since the pressure differentials are not large enough to invalidate the constant density assumption.

That is about it for fluids at rest. We are now ready to discuss fluids in motion. The first thing to remember is humility. Of all the problems in physics, accurately describing the motion of fluid is one of the most difficult. As always, our initial step is to classify things in order to focus our study on the simplest things first.

The main distinction to make is between turbulent and laminar flow. **Laminar flow** occurs when the streamlines, or flow of current, does not change over time. In a way, the fluid is both flowing and at rest since the overall pattern does not change though mass transport is occurring.

Turbulent flow is the opposite extreme and its precise description represents the last unsolved problem in classical Newtonian physics.[5] Frequently laminar and turbulent flow occur together (e.g., take a close look at the water that flows out of your tap at home). The initial flow is laminar but then breaks into turbulence after a certain point. Turbulence will often generate a vortex in its wake. The Reynolds number (which is related to viscosity to be discussed later) can be used to predict when turbulent flow will occur.

Even with laminar flow we can simplify things. We define an **ideal fluid** as both incompressible and non-viscous. By "non-viscous" (also known as **inviscid flow**) we mean we will ignore friction effects. When we say the fluid is **incompressible** we mean that the volume of any section of the fluid does not change as it flows. The shape of this section may stretch, twist, and turn but the total volume must remain the same.[6]

We can also describe incompressibility by saying that the volume flow rate is the same along the streamlines of the fluid. The **volume flow rate** is simply the amount of volume that flows through a given cross-section per second. It is pretty straightforward to see that the volume flow rate is equal to the product of the area of the cross-section and the speed of the fluid:

$$Q = \frac{\Delta V}{\Delta t} = \frac{A\Delta x}{\Delta t} = Av$$

So another (more useful) way to characterize incompressibility is by saying that

$$Q = A_1 v_1 = A_2 v_2 \tag{14.4}$$

for any two points along the fluid flow.

Now even if the fluid is compressible, the **mass flow rate** remains constant. In other words, if the volume increases the density will decrease because the mass of the fluid doesn't change. By using the logic in the previous paragraph we can show that

$$\rho_1 A_1 v_1 = \rho_2 A_2 v_2 \tag{14.5}$$

for any two points along the fluid flow. This equation is called the **continuity equation** since it represents the fact that the fluid is continuous: what goes in one end must come out the other. This also shows that a third way to characterize incompressibility is by saying the fluid density does not change along the streamlines.

When an incompressible fluid flows, it can be shown that

$$P + \tfrac{1}{2}\rho v^2 + \rho g z = \text{constant} \tag{14.6}$$

[4] One might call this the "pneumostatic equation".

[5] See the here for more info.

[6] We will also assume that the flow of the fluid is far below the speed of sound in the fluid. When the speed of sound is approached new behaviors are exhibited like shock waves and sonic booms.

where z is the elevation of the fluid. This follows from definitions of work and energy applied to the fluid. The conservation of this quantity is called **Bernoulli's equation** and can be re-written is several different ways. Notice that the hydrostatic equation (14.1) follows as a special case where $v = 0$.[7]

The term $\frac{1}{2}\rho v^2$ is sometimes called the **dynamic pressure** and can be adjusted for compressible flow (this adjustment is important when the speed approaches the speed of sound in the fluid). When the elevation differences are negligible, this dynamic pressure will drive the pressure differences in the fluid and is responsible for the **Venturi effect**. The drop in pressure that occurs when a fluid flows has several applications including the carburetor in your car and the force of lift on an airplane wing.[8] It's also why tarps get pulled of off moving pick-ups and why windows are blown out and roofs are torn off of buildings in high winds.

In applications where the elevation is important Bernoulli's equation (14.6) is often rewritten as

$$\frac{P}{\rho g} + \frac{v^2}{2g} + z = \text{constant}$$

The term $v^2/2g$ is called the **velocity head** of the fluid and the remainder is called the **hydraulic head** which is composed of the **pressure head** $P/\rho g$ and **elevation head** z. These terms and ideas are used in many applications such as geology and hydraulics engineering.

Incorporating corrections for the compressibility of fluids is quite challenging (and really requires a bit more understanding of thermodynamics), but we can talk a bit about viscosity. Ultimately viscosity is the friction between the streamlines of the fluid flow and give the fluid the ability to support shear stress. If we constrain a fluid between to large parallel plates and keep the bottom one at rest, the force required to maintain a fluid velocity v from the top place is given by

$$F = \eta A v d \tag{14.7}$$

where A is the area of the plates and d is the distance between them. The proportionality constant η represents the viscosity of the fluid.

If this viscosity is constant then we have a **Newtonian fluid**. Water and most gases are Newtonian fluids. One easy (and fun) example of a non-Newtonian fluid is cornstarch mixed in water (2:1 ratio works well). When placed under a large stress, the fluid becomes very viscous (will support a normal force like a solid), but when placed under a small stress acts like a normal liquid.

Stokes' law is a formula closely related to (14.7) and describes the force required to push a sphere with radius R through a fluid at velocity v:

$$F = 6\pi\eta R v \tag{14.8}$$

This formula can be used to estimate the air drag and an object when the velocity is small.

Poiseuille's law is another viscosity related formula. We have

$$Q = \frac{\pi R^4(\Delta P)}{8\eta L} \tag{14.9}$$

This formula describes the pressure differential required to produce a particular volume flow rate Q through a tube of radius R and length L. The higher the viscosity, the lower the volume flow rate.

Both Stokes' law and Poiseuille's law are all special cases of the use of viscosity in specific situations. They cannot be derived from Bernoulli's equation. They involve deriving the streamline pattern for the situation from more fundamental equations. As Bernoulli's equation is related to energy, we seek equations more like Newton's laws for the fluid. These are the **Navier-Stokes equations**. In

[7] Be aware that the h is the hydrostatic equation is actually the depth in the fluid, so $h = -\Delta z$.

[8] It is sometimes said that the air flows faster over the top of an airplane wing because the air that is split has a larger distance to cover in order to avoid being "cut in two". But the reality is that two streamlines of air do not rejoin in the same "place" (see the black line of dots in this animation.) The reality is that the air is compressed under the wing and this increases the speed of the air according to equation (14.4).

general these equations are a set of nonlinear partial differential equations for the velocity of the fluid. These kind of problems are among the most difficult in mathematics.[9]

However, it is possible to model these equations in the computer.[10] The computer will never give us the exact solution or a simple equation like those above. But the method is fairly straight-forward and works when other methods don't.

In fact, this is exactly how some CGI effects work. Many models are built to investigate the optimal design for cars, aircraft, even the Space Shuttle. This is also one of the only methods available to study turbulence.

In fact, it was through the computer simulation of a highly simplified model of the atmosphere that **deterministic chaos** was first (accidentally) discovered. Lorenz was running a computer simulation with his model and output the data. The next day he used the print out to restart the calculation and ended up with a completely different result. Ultimately the reason was that the accuracy of the print-out was to three digits, but the computer was calculating with six digits. The mathematical model was highly sensitive to the initial condition—a change in the fourth digit completely altered the simulation.

This extreme sensitivity to initial conditions is the hallmark of deterministic chaos. It is important to recognize that the apparent chaos in the system is exactly determined by the initial conditions—there is no random element. But the system is very sensitive to the starting conditions. This is also sometimes called the **butterfly effect** and explains why it is impossible to predict the weather with accuracy more than about a week into the future.[11]

An explosion is also extremely sensitive to initial conditions. The final location of the shrapnel depends on the precise way the explosive is constructed. But in order for a system to be chaotic the dynamics of the system must also be attracted back to its original state. The description of these **strange attractors** is one of the objectives in chaos theory. So any chaotic system must be both explosively sensitive and must "fold" back on itself in some sense.

Next week we will have a complete change of topic to discuss heat and temperature. This will occupy us for the next three lectures. We start by simply asking the question: how exactly does one measure temperature? Heat (typically) causes expansion and we will discuss the equations for that. In addition, there are three ways in which heat moves based on the phase of the object transferring the heat. We will also see that heat can change the phase of matter: from solid to liquid to gas, which will give us a physical distinction between heat and temperature. Finally we will round off the topic by discussing temperature extremes: the very hot and the very very cold.

[9] The equations in general relativity have this same character.

[10] For more info, see here.

[11] Not just difficult: mathematically impossible in principle. It is easier to control the weather (which is ultimately an engineering problem) than to predict it!

Lecture 15

Heat and Temperature

Read sections 12.1–12.9 and 13.1–13.4

What is heat? This is a question that goes way back. In antiquity, the ancient Greeks felt that the elements of the world could be categorized by two principles: wetness and heat. These principles combine in each of the four elements (earth, water, air, fire) in a rubric something like Figure 15.1.[1]

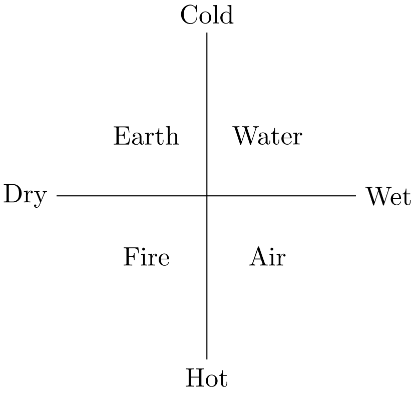


Figure 15.1: Periodic table of the ancient Greeks

Although we don't view the elements this way (we see in these ideas a foreshadowing of the phases of matter), the nature of heat was a subject of debate even into the 18th century. At that time there were three competing theories:

- Phlogiston: the fire-like element of the Greeks
- Caloric: a weightless, invisible fluid
- Kinetic: random motion of molecules

Ultimately the theory of phlogiston was rejected because it would have been required negative mass in order to remain consistent with the more accurate chemical experiments of the 19th century. However, both the caloric and kinetic theories of heat were viable at that time. Ultimately, the theory of a caloric fluid was subsumed as the conservation of energy and the kinetic theory is what we are all taught in grade school today. We will discuss the kinetic theory in more detail in Lecture 16.

Regardless of the conceptual model we keep with regard to heat, job one of any scientific approach is to measure it. For that purpose, we need to identify a **thermometric property**, that is, some physical effect that is controllably changed by temperature levels.[2] There many examples of thermometric properties:

- The volume of a liquid
- The length of a metal rod
- The pressure of a constant volume of gas
- The electrical resistance of a piece of wire
- The speed of sound
- The color of a hot stove

All of these properties and more are potential ways to quantify temperature. The first is the principle behind the standard glass thermometer (using either mercury or alcohol).

The SI unit for temperature is the Kelvin. We need some sort of agreed upon scale to measure temperature. It's okay to be looser when asking the question, "which

[1] Interestingly, most ancient civilizations had a rubric similar to this. See here.

[2] We will soon see a reason to create a distinction between the terms heat and temperature. For now I am switching to use the more proper term temperature for the current discussion.

is hotter," but to ask whether this is twice as hot is another problem altogether. The first temperature scale worthy of the name was developed by Fahrenheit and is roughly calibrated to the weather: zero is a cold day and one hundred degrees is a hot one. This **Fahrenheit scale** has largely been supplanted by the Celsius scale, but is still used by a small number of countries still behind the times.

Shortly after Fahrenheit, Celsius developed a more "metric" scale. This scale uses water as its basis. The freezing point of water is defined as zero degrees, and the boiling point of water is one hundred.[3] This **Celsius scale** is very easy to reproduce (by design) and is commonly used in labs today.

Fast forward a century and we have the **Kelvin scale**. Which is simply the Celsius scale calibrated to **absolute zero**. Using extrapolations based on the temperature dependence of the expansion of gases, this lower bound of temperature was discovered. There are several ways to interpret absolute zero, but perhaps the simplest is that this is the temperature at which all the random thermal motion in an object stops. This lower bound on the Celsius scale is $-273.15°$C. The Kelvin simply adds this number to the Celsius scale yielding "absolute zero" Kelvin.

The conversion between these scales is as follows:

$$F = \tfrac{9}{5}C + 32$$
$$C = \tfrac{5}{9}(F - 32)$$

and

$$C = K - 273.15$$
$$K = C + 273.15$$

One important thing to note is that the "size" of a degree in temperature is the same in Celsius and Kelvin. In other words, any temperature difference measured on the Kelvin scale is the same number as the difference on the Celsius scale. Frequently the important quantity in a problem is a difference in temperature. In that case we are indifferent to whether the temperature is measured in Celsius or Kelvin. However, there are some cases (e.g., the ideal gas law (16.1)) where the absolute temperature is required. In that case we must remember to convert Celsius into Kelvin before using our equations.

Any thermometric property can be represented mathematically as a function of temperature. As we discussed in Lecture 1 in equation (1.2), any function can be considered linear for small increments. Perhaps the simplest example of this is **thermal expansion**. Nearly all substances expand when their temperature increases (all else equal). We can write this as

$$\frac{\Delta L}{L_0} = \alpha \Delta T \tag{15.1}$$

where α is called the **coefficient of thermal expansion** which is a characteristic of the material.[4] Notice that the left-hand side of the equation is the same percent strain we introduced in Lecture 13. This thermal expansion accumulates over the length of the object. So even though the expansion may be a very small percentage, the total expansion may not be trivial. A few centimeters of expansion can slowly rip apart a highway or a battleship.

Usually α is positive indicating that things tend to expand with increasing temperature and shrink with decreasing temperature. One important exception is cold water. When water gets within four degrees Celsius of its freezing point, it will expand as the temperature drops. This has the effect of decreasing the density of the water. Generally, hot air rises because its expansion results in a lower density—it literally floats over cold air according to Archimedes' principle (14.2). By the same reasoning very cold water will float too. The result is that water freezes top down and ice cubes float. This also has the fringe benefit that

[3] These hundred degrees lead to this scale sometime being called the centigrade scale. This terminology is antiquated and a bit confusing as it seems to imply a more fundamental unit called a "grade" similar to the centimeter.

[4] There is also a version of this for volume expansion sometimes denoted β. This is usually used for liquids since one does not usually have a rod of fluid (though that is what a standard thermometer is). It's not too difficult to show that $\beta = 3\alpha$. This follows from calculating the new volume if each side expands according to equation (15.1) then applying the binomial theorem (1.3).

life can survive in a frozen-over lake because the ice acts as thermal insulation for the liquid water below. One common downside is that frozen pipes burst for the same reason. In fact, the freezing of water in rocks is a major contributor to geological erosion.

Having now established a way to quantify and measure temperature, we are prepared to discuss the flow of heat. Here is where we begin to create a distinction between the ideas of temperature and heat. We see temperature differences as the driver of heat flow.[5] For any solid heat will flow via **heat conduction**. The rate at which heat energy is transported (i.e., power) is given by

$$P = \frac{Q}{t} = k\frac{A}{L}\Delta T \tag{15.2}$$

where k is the **thermal conductivity** of the material. The larger the surface area A, the smaller the length L, the larger the temperature difference ΔT, the easier heat will flow.[6]

This formula is similar in form to other equations describing flow like Fick's law (16.5) and Ohm's law (24.3). The opposite of conduction is resistance. One quantity you may have seen before is the "R-value" which is defined as L/k. This metric is used rate the quality of insulation. Since it involves the reciprocal of thermal conductivity, this is a kind of thermal resistance. The better the insulation, the higher the R-value. An R-value of 5.6 per inch (typical US home insulation) corresponds roughly to one in SI units.

Heat conduction also occurs in fluids, but **heat convection** completely washes out any effect. Convection is caused by the heat being carried with the fluid rather than through it. Since no mass transport occurs in solids, there is no convection possible, but with fluids it is the main method of heat transfer. Unfortunately, fluid flow is so complicated that there is no single formula to describe convection.

The third and final way in which heat moves is through **heat radiation**. This is heat transported through the electromagnetic field. This is the weakest form of heat flow, but will work in the vacuum of space and is how the sun warms the earth. The formula for heat radiation is

$$P = \frac{Q}{t} = e\sigma AT^4 \tag{15.3}$$

where e is called the **emissivity** and is a property of the material ranging from zero to one. The value of σ is 5.67×10^{-8} and is called the **Stefan-Boltzmann constant**. It is a property of the electromagnetic field.[7] The surface area of the radiating object is A and T is its absolute temperature in Kelvin.[8]

Heat will flow whenever there is a temperature differential. But this still does not completely explain what makes some things hot and others cold. As we pour heat into a substance different materials require more heat than others in order to generate the same increase in temperature. This is called the **heat capacity** of the material. In general, we have:

$$Q = cm\Delta T \tag{15.4}$$

where c is the heat capacity, m is the mass involved and Q is the total heat flow required for the ΔT change in temperature.

As we continue to pour heat into a solid object, it will eventually melt. While it is undergoing this phase change, the temperature remains constant. The heat is going somewhere, but not into an increase in temperature. If one were to plot the temperature as a function of the heat input, the graph would flat-line here. This a characteristic of every **phase transition**. During the transition the substance is a mixture of the two phases—each at the same temperature.[9] A typical graph might look like Figure 15.2.

[5] Although it is tempting to think of temperature as a kind of potential energy creating a force which drives the flow of heat, this approach doesn't work. Essentially this is the caloric fluid model of heat. There are two fundamental problems: (1) temperature has a natural lower bound, and (2) it is better to associate heat with energy rather than the temperature.

[6] It should be noted that this is for the simplistic case of heat flowing straight through a rod. The conduction of heat through more complex geometries will obey different equations.

[7] I don't see any reason to not also have heat radiation via the gravitational field. The corresponding constant would be minuscule compared to the electromagnetic field however.

[8] This is an example of one of the formula where it is important to use Kelvin rather than Celsius. The easy way to remember which is that ΔT can be either Kelvin or Celsius, but just T must be Kelvin.

[9] In fact, it is possible for all three phases to coexist. This will only occur at a particular combination of temperature and pressure and is called the **triple point** of the substance. In fact, the modern definition of the Kelvin scale uses the triple point of water rather than the melting and boiling points originally used.

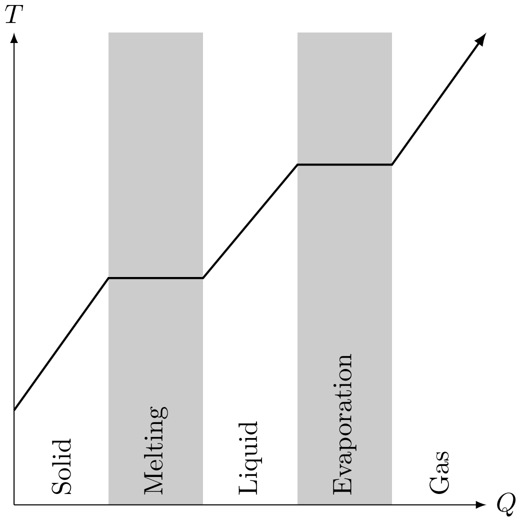


Figure 15.2: Typical phase change diagram

The slope on this graph represents the heat capacity of the substance. The horizontal distance during the phase changes represent the **latent heat** required to complete the phase transition. As such, each phase transition requires a certain amount of heat given by

$$Q = mL \tag{15.5}$$

Each transition has its own L value. This makes it possible for us to solve a basic calorimetry problem like the following.

Sometimes I make my coffee too hot, so I mix in a bit of ice to cool it down. Suppose I have 300 grams of coffee (which is basically water) at 80°C and mix it with 30 grams of ice at -10°C. What is the final temperature of the mixture?

Let's call the final temperature for which we are solving T. The amount of heat that flows out of the coffee is given by

$$Q_1 = (4186)(0.300)(T - 80)$$

because water has a specific heat of 4186. This heat flow will be negative because the coffee is losing heat.

The process for the ice is three-fold. First, its temperature needs to be brought up the boiling point:

$$(2000)(0.030)(10) = 600$$

then it must melt:

$$(335000)(0.030) = 10050$$

and then this melted ice must be brought to the final temperature T:

$$(4186)(0.030)(T - 0)$$

So the total heat that the ice absorbs is

$$Q_2 = 10650 + (125.58)(T)$$

Since the all the heat lost from the coffee is gained by the ice (no "leakage" of heat), the total change in heat is zero. Thus

$$Q_1 + Q_2 = 0$$

or

$$(4186)(0.300)(T - 80) + 10650 + (125.58)(T) = 0$$

Solving for T yields a final temperature of 65 degrees Celsius.

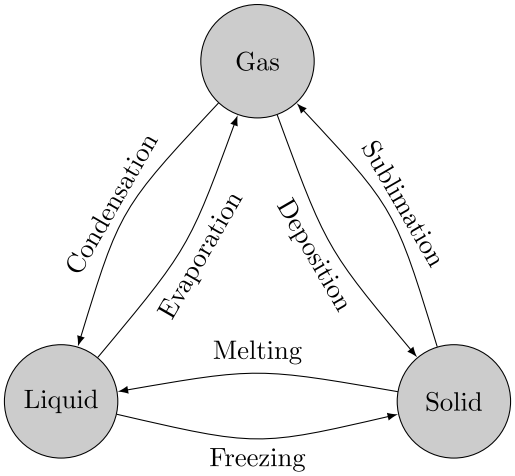


Figure 15.3: The six possible phase transitions between three phases

For specific combinations of pressure and volume, transitions between any two phases is possible. Consider the diagram in Figure 15.3.

In fact, it is possible to have even more phase transitions. As an extreme example, plutonium has six different solid phases making it very difficult to work with.[10] Liquid helium has two different phases with a transition at 2.17 kelvin. Below that value it acts as a super-fluid with zero viscosity (something like how a superconductor drops to zero resistance below a certain temperature—see below).

Finally, we will wrap up the lecture with a few notes on temperature extremes. At the hot end of the scale there is actually a fourth state of matter called **plasma**. This occurs when the random collisions between the atoms in a gas become so violent that they literally tear the atoms apart into ions and electrons. Plasmas are used frequently in high-tech manufacturing processes and other industrial applications.

But because plasma is electrically active, the fluid dynamics are extremely difficult to model. The currents of plasma have a tendency to curl and twist around one another and "pinch" in ways that make it difficult to control. This difficulty is one of the main obstacles in the creation of nuclear fusion energy.

On the other extreme is the very cold. One of the ways to cool down a gas is by letting it quickly expand—essentially reversing equation (15.1). In this way one can create temperatures cold enough to liquefy air. This makes liquid nitrogen pretty easy to acquire for practical uses. Obviously its extremely cold temperature can make it dangerous if used improperly.

Getting closer and closer to absolute zero is quite a story in the history of physics. Each increment is more and more difficult to accomplish (like approaching the speed of light). Along the way we find **superconductivity**. First discovered in the early 1900s, this phenomena is when the resistance of certain metals drops to zero—not close to, but literally zero. This is one of the few quantum mechanical effects on a macroscopic scale (the laser is another). As temperatures drop, these quantum mechanical effects become more evident in various ways. Superconductivity is one of the most dramatic examples of this.

Next week we will continue the story of heat and temperature. The focus will be on kinetic theory. By assuming that molecules in a gas interact via simple elastic collisions, we will be able to derive the relationship between pressure, volume, and temperature for an ideal gas. This approach is called statistical mechanics and involves quite a bit of math, so we will only be able to touch the surface of this subject. We will also see that diffusion can be explained using the same kinetic theory. In the end we will see that classical statistical mechanics is not as successful as it ought to be given the success of Newtonian mechanics. These

[10] See here for more details.

failures pave the way to the discovery of quantum mechanics which we will review in Lecture 22.

Lecture 16

Kinetic Theory

Read sections 14.1–14.4, review Lecture 8

In the previous lecture we investigated some of the basic properties of heat and temperature by focusing on what heat "does": how to measure it, how it flows, etc. In this lecture we focus on what heat "is": the random motion of molecules.

Richard Feynman has said,

> If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words? I believe it is the **atomic hypothesis** that
>
> All things are made of atoms—little particles that that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.
>
> In that one sentence, you will see, there is an enormous amount of information about the world, if just a little imagination and thinking are applied.

Well, today we will apply this idea in the context of the kinetic theory of gases. The basic idea is that the higher the temperature of a gas, the more kinetic energy the molecules possess. We can even incorporate phase changes into this framework by associating latent heat with the potential energy created through inter-molecular interactions.[1]

Before we can speak in detail about molecular motion we need to deal with their size. The radius of a typical atom is on the order of the nanometer. This means that the number of atoms in any laboratory sample is huge. It also means that they are essentially unobservable without highly specialized equipment. This was a major obstacle for the adoption of the atomic hypothesis in the early 1900s. It was none other than Albert Einstein who analyzed **Brownian motion** (the random, jerky motion of particulates suspended in a fluid) that provided the first indisputable physical evidence for the atom.

The point here is that the fundamental unit of the **mole**[2] has been developed to deal with this atomic size. By definition, the mole is the number of atoms in 12 grams of carbon-12 (a specific isotope of carbon—see Lecture 29) and has a value of 6.022×10^{23}. This number is called **Avagadro's number** labeled N_A.

You may wonder why we need a special unit for something that is just a number. Isn't it just a naming convention, like a dozen eggs? In a fundamental sense that is true, but one must remember that we are human. No matter how sophisticated our machinery becomes, ultimately measurements must be translated to and from

[1] A truly phenomenal website with a simple molecular dynamics applet is located here.

[2] Abbreviation: mol. I believe that this is simply a shortening of the word molecule.

this human scale. There is no way to literally count the number of atoms in a particular sample.[3] In other words, the mole is not just a number that we simply arbitrarily choose—it represents the relationship between the lab and the atomic world. This relationship is best summarized in the following formula:

$$\frac{\text{grams}}{\text{mole}} = \frac{\text{molecular mass}}{\text{molecule}}$$

where the **molecular mass** is the mass of each molecule in so called **atomic mass units**. By definition a carbon-12 atom has a mass of 12 amu. The atomic mass units are listed in any periodic chart, so given the chemical make-up of a substance we can determine its total molecular mass and therefore the number of moles in a particular sample.

With this link back to the lab we are now prepared to dive into the world of the microscopic. The simplest microscopic system to analyze is an **ideal gas**. In an ideal gas, we assume that the molecules have two properties.

- Negligible interaction (which will eliminate any possible phase transitions)
- Negligible size (which will allow the volume to decrease without limit)

We relax both of these conditions enough to allow the molecules to interact through elastic collisions. These collisions will be the only mechanism to distribute kinetic energy throughout the gas (other than collision with the walls of the container). Most gases at standard temperature and pressure[4] can be considered ideal.

As mentioned in the previous lecture, gases will expand with increasing temperature. This volume is obviously also related to the pressure surrounding the gas. These three variables combine in the **ideal gas law**

$$PV = nRT \tag{16.1}$$

where n is the number of moles of gas and R is a proportionality constant equal to 8.31 in SI units. An alternate form is

$$PV = NkT \tag{16.2}$$

where N represents the number of molecules in the gas and $k = R/N_A$ is called the **Boltzmann constant**.

Now we are ready to introduce some mechanics. Consider a single particle trapped in a cubical box of length L. Suppose the particle moves parallel with one of the edges of the box—we will call this direction $\hat{x}$. Assuming the atom strikes and rebounds off the way elastically (this is an ideal gas), the change in momentum is given by

$$\Delta p = 2mv$$

This collision will occur every time the atom travels the length of the box twice. Given the velocity v, we have

$$\Delta t = 2L/v$$

Combining these two results according to the impulse form of Newton's law (8.2) gives us an average force per collision[5] of

$$\langle F \rangle = mv^2/L$$

Since pressure is average force over area, we may write

$$PV = mv^2$$

The right-hand side is simply twice the kinetic energy of the atom.

[3] Maybe with the advances in nanotechnology this will be possible. If that day ever comes, we will still need to define the counting process in a way that is reproducible and independent of external variables.

[4] Most commonly 0°C and 101.3 kPa (atmospheric pressure), though some organizations use different definitions.

[5] This calculation is similar to the "machine gun pressure" problem in Lecture 8 on page 44.

Now if we consider a random collection of molecules, the momentum will be split into three due to the three dimension of space, so in general we have

$$PV = \tfrac{2}{3}N\langle KE\rangle$$

where $\langle KE\rangle$ represents the average kinetic energy per molecule. Comparing this with 16.2 yields

$$\langle KE\rangle = \tfrac{3}{2}kT \tag{16.3}$$

which explicitly shows the relationship between the temperature of the gas and the random motion of its molecules. This average motion is called the **internal energy** of the gas. Its the kinetic energy of the system after the bulk motion has been isolated. We have shown the internal energy of an ideal gas to be $\frac{3}{2}NkT$ or $\frac{3}{2}nRT$.

The above derivation of the internal energy of an ideal gas is an example of the use of **statistical mechanics**. Essentially we assume that our system is composed of a large number of identical pieces (atoms). By applying the laws of mechanics to each part, we can take the overall average to derive the properties of the overall system. For example, if we relax our restrictions on an ideal gas a bit we can derive the **Van der Walls equation**:

$$\left(P + a\frac{n^2}{V^2}\right)(V - nb) = nRT \tag{16.4}$$

where a is related to the attraction between the molecules and b the "size" of the molecules. This represents an improvement over the ideal gas law (16.1) and is applicable to liquids as well.

Derivations can go the other way too. The **Maxwell-Boltzmann distribution** represents the number of molecules at various speeds in the gas. There are several simulators on the web to show how this distribution works (here's one: hit setup then go, use the slider to slow down the animation). The point is that even if the molecules all start with the same speed, through random collisions some will increase in speed and some will decrease. The distribution of speeds over the long haul is what the Maxwell-Boltzmann distribution describes.

An example of the distribution is given in Figure 16.1. One thing to note is that although the average speed does increases with temperature, there is always a range to the speeds: there are always some that are slow and some that are fast. This is one of the ways to explain evaporation. No matter what the overall temperature of the liquid, there will always be some molecules with high kinetic energies. These molecules have enough energy to escape the potential well that binds the liquid together and escape.

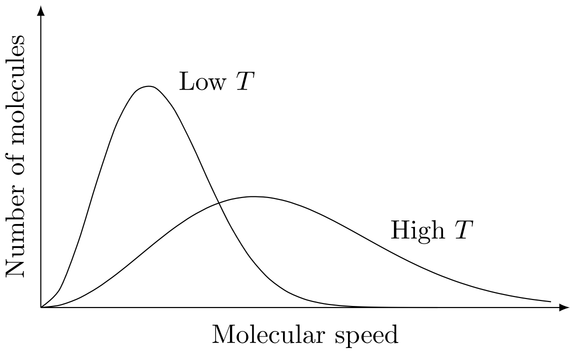


Figure 16.1: Example of the Maxwell-Boltzmann distribution of speeds in an ideal gas.

On the technical side, notice that the distribution is not symmetric (no negative values). This implies that the most probable speed (the peak of the graph) is not the same as the average speed. It can be shown that the average is about 13% larger than the peak value.

In fact, the distribution of the kinetic energies depends on v^2, so the shape is different though qualitatively the same. But the average kinetic energy does not correspond to either the most probable speed or the average speed. The speed to which it does correspond is called the **root-mean-square** (abbreviated as RMS) because it is the square root of the average of the square of the speeds. This complicated sounding calculation is actually quite common and shows up in a variety of contexts related to statistics. Usually called the **standard deviation**[6], it is a good measure of the distribution in the variable of interest. One way to see this is that the straight average will cancel negative and positive contributions (which gets you to the middle), but the root-mean-square will accumulate them (which gets at the spread of values). For the Maxwell-Boltzmann distribution, the RMS speed is about 22% larger than the peak value.

Diffusion is the process in which solutes flow from regions of high concentration to low concentration. The substrate through which the flow occurs could be solid, liquid or gaseous. Diffusion is comparable to heat conduction in many ways and is easy to understand using kinetic theory. Some of the molecules on the boundary of the high concentration region will have a random motion away from that region. As these molecules flow away they reach the area of low concentration. Though molecules are leaving the low concentration area also, they do so at a smaller rate. More flow in than out. Eventually a concentration equilibrium is established and there is not net flow of material (though the individual molecules are constantly moving back and forth).

The equation governing diffusion is **Fick's law**:

$$\frac{M}{t} = D\frac{A}{L}\Delta C \tag{16.5}$$

Notice the similarity to the equation for thermal conduction (15.2). The amount of mass M transported per second is proportional to the concentration differential ΔC, the geometry of the situation A/L, and a proportionality constant D called the **diffusion constant**.

This is the basic process behind **osmosis** which has so many applications in biology. The osmotic pressure due to the concentration differentials in water is responsible for the rigidity of plants, the stability of the cell, and is how plants draw water and nutrients from the soil.

Diffusion is also used in many high-tech processes to control the electrical properties of various substances. We will talk about some of these applications in Lecture 28.

It is easy to take the success of kinetic theory for granted. For over two centuries the applications of Newtonian mechanics continued to spread until some began to think there was no other way the universe could work. They were wrong and the first signs were discovered in the decades prior to and after 1900. Of the two scientific "revolutions" around this time, the changes wrought by quantum mechanics exceed that of relativity. When viewed the right way, relativity can be seen as a tweak to the framework of Newtonian mechanics—quantum mechanics reworks the very conceptual foundation. We will return to this topic in Lecture 22 and discuss relativity and quantum mechanics in more detail in Lectures 26 and 27. But for now it is worth mentioning some of the cracks that were discovered in this foundation related to kinetic theory.

The first crack we have already mentioned: superconductivity. As temperatures get down near absolute zero, the electrical resistance of certain materials drop to zero. We shouldn't be surprised to find that electrical resistance drops with decreasing temperature, but exactly zero implies new physics.

The second crack is the heat capacity of solids. According to kinetic theory the heat capacity of any solid should be $3R$. This is based on a result from statistical mechanics called the **equipartition principle**. It states that the internal energy

[6] Actually, for technical reasons the definition of standard deviation is not quite the same as root-mean-square.

of a system in thermal equilibrium is $\frac{1}{2}kT$ for each degree of freedom. We have seen an example of this in equation (16.3)—the particle has three degrees of freedom corresponding to the three dimensions of motion. In fact, the internal energy of a diatomic molecule is $\frac{7}{2}kT$ per molecule because it has seven degrees of freedom.[7]

For atoms locked in a solid, there is no translation or rotation, but they can vibrate in three dimensions which means that there are six degrees of freedom per molecule, so the internal energy of the solid ought to be $3NkT$ which corresponds to a specific heat capacity of $3R$. However, as temperatures get cold the heat capacity of a real solid drops. It is as if some of the degrees of freedom are getting "frozen" out.[8] This is purely quantum mechanical effect. Einstein created the first quantum model that correctly predicted this effect.

The third crack involves something called the **ultraviolet catastrophe**. Historically, this is the start of all things quantum. This problem is related to the heat radiation from a so-called **blackbody**, or an object with emissivity equal to one. The energy spectrum from a black body is qualitatively like that of the distribution in Figure 16.1,[9] but classical statistical mechanics predicts the spectrum to extend into the low wavelength (ultraviolet) without limit. This would imply that all radiation should occur beyond the visible spectrum which anyone who has sat in front of a campfire would say is clearly nonsense. Interestingly, Einstein also had a role to play in recognizing the way that the quantum hypothesis (introduced by Planck) could resolve this embarrassment.

Next week we will finish off our discussion of heat and temperature by talking about thermodynamics proper. This can be seen as a continuation of Lecture 15, but we will need our calculations of internal energy (16.3) to complete the picture. We will state and discuss the laws of thermodynamics, the first of which is simply the conservation of energy. Using simplified thermodynamic processes we will talk about efficiency in heat engines which will lead us to discuss reversibility. This in turn will motivate the definition of a new quantity called entropy which is the subject of the (infamous) second law of thermodynamics.

[7] Three for translation, two for rotation, and two for vibration. There are only two for translation because rotation about the axis that connects the two atoms is no motion at all (symmetry). Two for vibration because the energy has both kinetic and potential modes.

[8] In fact, a typical diatomic gas (like nitrogen or oxygen) actually has a specific heat capacity closer to 2.5 because the vibrational modes are inactive. More complex molecules are locked down to about 3. See here.

[9] I'm just too lazy to draw another diagram—the equations are different. See Figure 22.1 for an accurate picture.

Lecture 17

Thermodynamics

Read sections 15.3–15.12

The science of thermodynamics is a product of the industrial revolution and the pursuit of a more efficient steam engine. As such the approach has a strong engineering emphasis—practical rather than theoretical. Another way in which it is distinguished from mechanics is that the approach has more "systems" reasoning: we focus on how the properties of the systems interact and change. Though statistical mechanics sheds much light on thermodynamics, technically they are independent. As such, the laws of thermodynamics are is still valid when quantum mechanics replaces classical mechanics.

Our starting point is the conservation of energy. This is called the **first law of thermodynamics** and is usually written as

$$\Delta U = Q - W \tag{17.1}$$

Basically this is saying two things:

- Heating a system increases its internal energy ($\Delta U = Q$).
- Work done by a system decreases its internal energy ($\Delta U = -W$).

Equation (17.1) is written in the context of heat engines, so the concern is how much work we can get out of the system—which is why there is a negative sign in front of W.

Remember that the internal energy of a system is directly related to its temperature. For an ideal gas we have

$$\Delta U = dkT/2$$

where d is the number of degrees of freedom for the molecule—we will assume $d = 3$ for a monatomic gas (like argon) and $d = 5$ for a diatomic gas (like oxygen) (see page **??**).

The first law doesn't seem to say much until we start to add some information about the system under consideration. The simplest system to consider is an ideal gas, for which we have the ideal gas law (16.1). There are four basic parameters involved: pressure, volume, amount of material (moles), and temperature. The first two are related to work through $W = P\Delta V$.[1]

We will normally assume the amount of material is constant in a particular process. However, the other three variables may change in various ways. We will focus on four simpler processes each of which hold a particular variable constant.

[1] This follows from the definition of work $W = F\Delta x$. Divide the first factor by A to get pressure and multiply it into the second factor to get V.

The first process is **isothermal**, meaning that the temperature is held constant. A common way to do this is to connect the system to a **heat reservoir**.[2] This is simply an outside system which is considered so large that the change in its temperature is negligible due to the flow of heat to and from our system of interest.

Since the temperature is held constant, we know two things. First, the change in internal energy is zero. Second, from the ideal gas law we know that the product of pressure and volume is constant. Usually in an engine we manipulate the volume (driving a piston, for example), so we are interested in the work done when the volume changes in the gas. The challenge here is that the pressure changes too. This calculation involves a bit of calculus, so I will simply quote the result:

$$W_{\mathrm{isoth}} = nRT\ln(V/V_0) \tag{17.2}$$

A volume increase corresponds to positive work. Since the temperature is held constant, heat must be flowing into the system from the reservoir to support the process. The amount of heat is equal to the work done according to the first law. This isothermal process shows that it is possible to manipulate the flow of heat without a temperature differential.

The second process to consider is **isobaric**, meaning that the pressure is held constant. In this case, the ideal gas law tells us that ratio of volume and temperature is constant. The work done is simply $W_{\mathrm{isoba}} = P\Delta V$ because the pressure is constant. An easy way to support this kind of process is through an open container exposed to the atmosphere.

The third process is one with constant volume, like with a closed container. This is called a **isochoric** process. There is no work involved since $\Delta V = 0$.

The fourth and final process we will consider is called **adiabatic** which means no heat flow. A thermally insulated container will obviously be adiabatic, but also any process that happens quickly. If the process is fast enough, there simply is not enough time for the heat to flow. So, whether a given process is adiabatic or not depends on the thermal conductivity of the materials. But if there is no heat flow, the first law reduces to $\Delta U = -W$, so the work done in an adiabatic process comes from the internal energy of the gas (and a corresponding decrease in temperature). We have:

$$W_{\mathrm{adiab}} = -\tfrac{3}{2}nR\Delta T \tag{17.3}$$

for a monatomic ideal gas.

In a way, the adiabatic process stands in contrast to the isothermal process. Both involve divorcing the connection between heat and temperature. In the adiabatic flow temperature changes without heat flow but in this isothermal process heat flows without a temperature change. The connection between pressure and volume is similar too. For an adiabatic process we have

$$PV^{\gamma} = \text{constant} \tag{17.4}$$

where γ is a number larger than one.[3] For an ideal gas, we have

$$\gamma = \frac{d+2}{d}$$

where d is the number of degrees of freedom. For the monatomic ideal gas, $\gamma = 5/3$. This shows that under compression the pressure in an adiabatic process will rise faster than in an isothermal one. Under expansion the adiabatic pressure will fall faster.

These four processes show us how it is possible to use the first law (17.1) to convert heat flow into usable mechanical work. Any contraption that harnesses heat flow to create work is called a **heat engine**. Thermodynamics is primarily

[2] Sometimes called a "heat bath".

[3] It can be shown that this γ is equal to the ratio of the specific heat capacity of the gas at constant pressure relative to that at constant volume.

the mathematics behind heat engines. Every heat engine is rated by its **efficiency** which is the power it generates in useful work divided by the power it consumes.[4] This is equivalent to the definition we used in Lecture 7 of total work done over the energy consumed because power is work per second.

Since a heat engine converts heat flow into work its primary waste product will be heat energy. Since energy is conserved, we have a simple relationship

$$Q_{\text{in}} = W + Q_{\text{out}}$$

which simply says that the energy going in either comes out as usable work or wasted heat. This allows us to write a formula for efficiency in terms of heat:

$$e = 1 - \frac{Q_{\text{out}}}{Q_{\text{in}}} \tag{17.5}$$

So efficiency is the opposite of how much energy is wasted per unit of input energy.

With much fanfare, we are now ready to introduce the **second law of thermodynamics**:

> Between high and low temperature, heat will flow spontaneously

Truly an inspiring statement, huh? I think the reason why there is so much talk about the second law is that this sentence is a Trojan horse of sorts. This simple, obvious statement is the reason why no heat engine will ever be close to 100% efficient, why no perpetual motion machine can exist, and can even explain the ultimate "heat death" of the universe. Understanding how this can be has lead to many alternate formulations of this law. We will discover a few along our way.

The key word in the previous statement is "spontaneously" which is closely related to **irreversibility**. Heat will not flow spontaneously from cold to hot. Of course, heat can flow against a temperature differential (this is what a refrigerator does after all). But when the heat flow is spontaneous, the system does not control the flow of energy. This is unlike a mechanical system in which the energy is flowing due to the internal forces between the parts. The mechanical energy distribution is directly related to the configuration of the system, so it is possible (in principle) to reverse the flow of energy by simply reverting this configuration back to its initial state. There is nothing a mechanical system can do to reverse the spontaneous flow of heat.

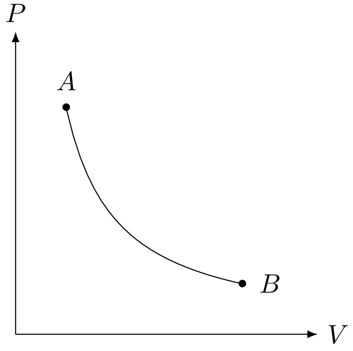


Figure 17.1: An isothermal process for an ideal gas plotted on PV chart

So spontaneous heat flow is bad for efficiency. We will now see that even a perfectly reversible heat engine cannot be 100% efficient. A heat engine requires the flow of heat. But the second law implies that only isothermal heat flow is reversible. In addition, any engine must cycle: the process must repeat. The simplest engine we can consider is one that uses an ideal gas (we've already calculated the work involved for all four basic processes). A complete cycle will involve returning to the same pressure, volume, and temperature. In fact, it's sufficient to track just pressure and volume since if they match, so will the temperature. We can plot the process on a PV-chart like Figure 17.1.

Simply using isothermal processes is not enough for our engine, however. We could create a cycle from A to B and back, for example, but the work done in the expansion process—given by equation (17.2)—is equal to the work required to compress the gas back to its initial state. We have an engine but it does no net work.

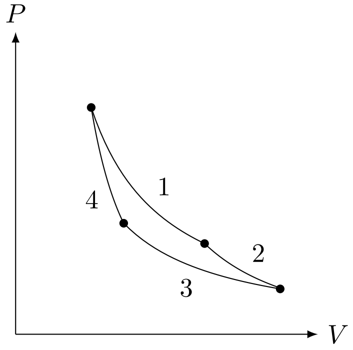


Figure 17.2: Example of the Carnot reversible cycle for a monatomic ideal gas

We can do better. An adiabatic process is also reversible because there is no heat flow at all. This will give us a way to change the temperature of the gas. The simplest non-trivial process which combines isothermal and adiabatic processes is called a **Carnot cycle** (see Figure 17.2).

[4] I like to draw a distinction between a machine which multiplies force, and an engine which does work. In my mind the key metric for a machine is its mechanical advantage, but for an engine it is power: how fast it creates energy. I try and keep a distinction in the nomenclature although I don't think most authors do.

Suppose we start the ideal gas in a state of high pressure and low volume. The first step is isothermal expansion. For example, we allow the gas to expand while thermally connected to a heat reservoir at temperature T_1. The work done by the gas is given by equation (17.2). The second step is adiabatic expansion. We now disconnect the gas from the heat reservoir, thermally isolate the gas, and allow it to continue to expand. The work the gas does is given by (17.3). The gas is now in a low pressure, high volume state with a temperature lower than the initial state.

This is all the work we are going to get out of the engine. We now need to work on returning the gas back to it original state (high pressure, low volume) in order to complete the cycle so we need to compress it. Adiabatic compression will simply reverse what we have already done, so the third step should be isothermal compression. Thermally connect the gas to another heat reservoir at temperature T_2 and compress it.[5] The fourth step is adiabatic compression which will bring the gas back to its original state.

Work is required in the last two steps, but it will be less than the work done by the gas.[6] We know that the efficiency of the engine is the net work done divided by the input heat energy. This input heat energy occurs during the isothermal expansion. The amount of heat is given by equation (17.2) because of the first law and the fact that the internal energy does not change.

The heat wasted occurs during the isothermal compression and is also given by (17.2). Putting all of these facts together gives us the efficiency for our engine:

$$e = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} \tag{17.6}$$

Here's the kicker (and also the insight that put Carnot in the history books). This is the best any heat engine can do between reservoirs of these two temperatures. The argument has two parts. The first is simply that any irreversible engine will have a lower efficiency than a reversible one. This is pretty obvious as any spontaneous heat flow will simply add to the wasted heat without any productive work. The second is that every reversible engine will have the same efficiency. One way to see this is by supposing we have two reversible engines with different efficiencies. Since they are truly reversible, run the more efficient one backward by using the work generated from the first one. In this way we generate excess work without any net heat flow at all!

So, although equation (17.6) was derived for an ideal gas, it represent the maximum possible efficiency of any heat engine. Obviously a real heat engine will be worse.

Take a moment to compare (17.6) with (17.5). They allow us to write

$$\frac{Q_{\text{hot}}}{T_{\text{hot}}} = \frac{Q_{\text{cold}}}{T_{\text{cold}}}$$

for the Carnot cycle we have been discussing. This suggests there is a kind of quality of heat flow at various temperatures. In a way, the quality of heat flow at a lower temperature is greater than at a higher temperature. This "quality" is related to reversibility and is not called **entropy**.

Entropy is a way of grading the internal energy of a system and is closely related to how closely the system is to thermal equilibrium. When the internal energy of a system is far from equilibrium, it is said to be in a low entropy state. For example, when one hot and one cold object are thermally connected (forming one system) heat will flow. In other words, the internal energy of the combined system redistributes itself until it reaches thermal equilibrium (uniform temperature). Another way to say this is that the system moves from a low to a high entropy state.

[5] How far should we compress it? It takes a bit of calculation, but we need $V_D/V_C = V_A/V_B$. This will cause the fourth step to actually complete the cycle.

[6] The net work is actually the area enclosed within the curve from Figure 17.2.

Whenever heat flows out of a system, the entropy changes by

$$\Delta S = \frac{Q}{T} \tag{17.7}$$

where S represents the entropy of the system.[7] By this definition, the entropy change in the Carnot cycle is zero. In this way the change in entropy represents the "quality" of the heat flow. In an irreversible cycle, the entropy created when the heat flows out of the system at the low temperature exceeds the reduction that occurs when heat is absorbed at the high temperature. The net entropy change is positive. This is a characteristic of any irreversible cycle.

Entropy is a recognition of the fact that internal energy tends to uniformity. In this way, the **second law of thermodynamics** is often written as

$$\Delta S \geq 0 \tag{17.8}$$

Be aware what this statement does not say: it does not say entropy never decreases—it often does. But whenever the internal energy of a system is pushed away from thermal equilibrium, the entropy increases in the external environment that is doing the pushing. At best the total entropy (system plus surroundings) is zero for a reversible process. The total entropy increases for an irreversible process. Every decrease in entropy is accompanied by a larger increase in entropy somewhere else.[8]

The heat flow that occurs into the cold reservoir is wasted energy and is lost for mechanical work. The entropy change of the process offers us a way to quantify this through the formula

$$W_{\text{lost}} = T_0 \Delta S \tag{17.9}$$

where T_0 is the temperature of the coldest reservoir in the system.

We are now prepared to come full circle and discuss entropy in the context of kinetic theory. I have already mentioned how entropy represents the distribution of the energy in a system. For example, if we take a system and divide into two parts, the energy distribution will follow the degrees of freedom in the system.

We define the microscopic state our system as the way in which the energy is distributed. It can be shown that the number of states available to a system is exponentially dependent upon its degrees of freedom:

$$\Omega \propto E^{d/2} \tag{17.10}$$

Since the degrees of freedom of a typical gas is related to the number of its atoms, this can become a very large number. The number of ways in which the energy can be distributed uniformly far outnumber the other ways it can be distributed.

A simple example calculation is shown in Table 17.1 for a system with five units of energy. The system has two parts with six and four degrees of freedom in the first and second parts, respectively. Notice how the largest number of total states occurs when the energy per degree of freedom is equal in the two parts. Increasing the degrees of freedom makes the distribution much more extreme.

E_1	E_2	Ω_1	Ω_2	Ω
0	5	0	25	0
1	4	1	16	16
2	3	8	9	72
3	2	27	4	108
4	1	64	1	64
5	0	125	0	0

Table 17.1: Calculation of the number of states in a simple two-part system

The basic postulate of kinetic theory in this context is that the probability of the system being in any microscopic state is equal. In effect this means that the system is usually in a state of uniform distribution. The system is constantly bouncing between different states, but there are so many that the effect is to pull the system to uniformity—which we have previously called equilibrium and "explains" the second law.

Boltzman was able to show that the absolute entropy of a system is related to the number of states by

$$S = k \ln \Omega \tag{17.11}$$

[7] This formula only works for an reversible isothermal process. In order to calculate the entropy change for an arbitrary process, it is sufficient to do so for another process that connects the beginning and end states. The entropy change is independent of the process used to get there.

[8] There is a small loophole here. Another way to affect the entropy of a system is by combining it with another: the total entropy is the sum of the two parts. If the absolute entropy of a system could ever be negative, we could reduce the total in this way. The so-called **third law of thermodynamics** closes this loophole by stating that the entropy of a system is only zero at absolute zero. In this way, the entropy of a system can never be made negative.

a result of which he was so proud that he has it engraved on his tombstone.

This completes our survey of thermodynamics. We will switch gears again to discuss wave motion. This will occupy us for the next couple of lectures and then we will apply some of these results to the study of sound and light. In the first lecture we will discuss how waves are generated from a single source. The primary effect is the radiation of energy through some medium whether it is a string, the air, or the electromagnetic field. The speed, wavelength, and intensity of the wave will be discussed. We will extend the topic a bit to include multiple sources at the same location. In the lecture after we will discuss multiple sources separated in space (which leads to interference—the chief characteristic of waves.).

Lecture 18

Wave Motion and Radiation

Read sections 16.1–16.9

Radiation can be divided into two classes: that which transports mass and that which transports energy. An example of the first class is nuclear radiation. When nuclear energy is released, each atomic nucleus acts like a little bomb. The atomic particles go flying in all directions carrying residual kinetic energy. This atomic shrapnel we call radiation. We will see in Lecture 29 that there are three types of natural nuclear radiation each with varying properties.

The second class of radiation is that which transports only energy. This is the kind we refer to when a pebble is dropped in a still pond and we say that the ripples radiate from the impact. This radiation does real work—each wave erodes the shoreline a tiny bit—so energy is truly being transported.

In the first case, the radiation is simply mass in motion. This mass carries momentum and energy which interacts with other matter through the principles we have already studied. But the second case is different. No mass moves from point A to point B, only energy. This energy transport requires a **wave medium**. The medium does not itself move from point A to point B, but it does support this transmission of energy.

In order to support this transmission of energy, the medium must be composed of interconnected parts.[1] There are many different types of media: strings are the simplest model, but we will also talk about the way waves move through solids and fluids. The electromagnetic field also acts as a medium for waves, although it is a bit of a special case (we will revisit this issue in Lecture 21 and 22).

Since the applications of wave motion are so wide spread, it is useful to keep the discussion fairly abstract. However, most wave phenomena can be seen in a simple string. The essential characteristic of the wave medium is that it be composed of coupled oscillators all near equilibrium.

Although the medium itself does not move in bulk, there is a direction involved in the radiation of energy. We say that we have a **transverse wave** if the oscillation in the medium is perpendicular to the movement of the disturbance (like a string). We have a **longitudinal wave** if the oscillation is parallel to the radiation (like a compressed spring, or sound waves). Wave motion is not confined to either/or: typical water waves have both transverse and longitudinal components.

By definition, the disturbance of a transverse wave has only one degree of freedom, but a longitudinal wave has two (the y and the z if it is moving along the x-direction). To completely describe the wave we need to specify its direction. This is called the **polarization** of the wave.[2] In addition, the disturbance need not be a physical displacement. We could speak of a heat wave, for example. As long

[1] This interconnection is the crucial ingredient and makes the wave motion possible—when one particle moves, those connected will be dragged along with it. These in turn will drag additional parts which propagates the disturbance through the entire medium.

[2] It is a bit more complicated than this since we could drive the string with a circular type of motion. In that case, the direction of the oscillation rotates in time: this is called circular polarization. More complex combinations are possible as well.

as the medium is coupled across this property, radiation will occur. The possible polarization states depend on the nature of the medium.

Because we are essentially discussing the motion of oscillators, the mathematics begins where Lecture 12 left off. The basic result was that when a system is oscillating sufficiently near equilibrium, the oscillation is governed by equation (12.5). However, our oscillators are connected to one another, and the radiation has a dampening effect as the energy is propagated away.[3] The oscillation of each element in the medium looks more like Figure 12.5.

Now instead of focusing on each element in the medium, we want to discuss the profile of the disturbance across the medium. For simplicity we will talk about an string with a linear mass density of μ under tension T. An element of the string will experience a net force when it and its two neighbors are not in a straight line (see Figure 18.1). The net force due to the curvature in the string acts to eliminate that curvature. In fact, the stronger the curvature in that "line", the greater force of restoration. Mathematically, when the forces of restoration are proportional to the curvature of the disturbance, the medium obeys the **wave equation** and will support the kinds of waves we are discussing in this lecture.

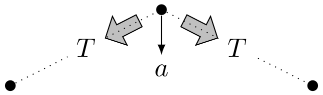


Figure 18.1: Forces on the element of a string eliminate curvature

The speed with which any disturbance propagates through this string is also related to the tension and linear mass density. For an ideal string, the speed is given by

$$v = \sqrt{T/\mu} \tag{18.1}$$

For a solid this speed represents the speed of sound in the substance. We have

$$v = \sqrt{Y/\rho} \tag{18.2}$$

Actually, this is only the formula for a compression wave in a solid. When a solid is disturbed, the strain can propagate through shear stress as well.[4] For a liquid we have

$$v = \sqrt{B/\rho} \tag{18.3}$$

where B is the bulk modulus. Since the pressure in a gas is temperature dependent, so is the speed of sound. We have

$$v = \sqrt{\gamma k T/m} \tag{18.4}$$

where m represents the molecular mass of the gas.[5]

If we pluck a string, the oscillation will eventually die away as the energy is radiated through the string (assuming the string is infinite in length). In order to maintain a sustained wave pattern, we must have a source that supplies the energy that is being drained away. The simplest driver is simply one that moves with simple harmonic motion. As we discussed in Lecture 12, when an oscillator is driven by an outside source, its frequency will match the frequency of the driver regardless of its natural frequency. In this way we can control the motion of the medium from this single driver: all the other elements of the medium (that participate in the radiation) will vibrate with this same frequency.

If we consider a string that is being driven at one end by a simple harmonic oscillator, then the resulting disturbance will look like Figure 18.2. In this figure the wave is traveling to the right. Notice how each element of the string oscillates with the same frequency (the white dots, for example) as the driver on the far left. If the motion of the element matches the driver, it is said to be **in phase** (the gray dots are in phase). If the motion is the exact opposite (when one is up the other is down—like the white dot), it is said to be **out of phase**.

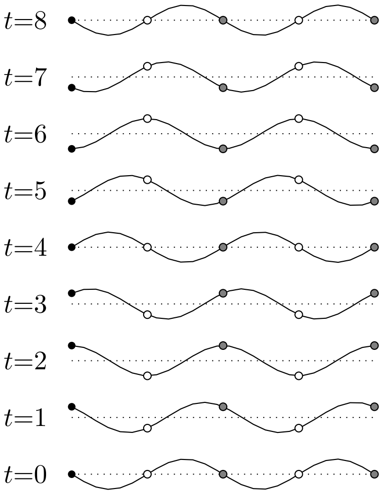


Figure 18.2: Sine wave profile over time

It is no coincidence that we use same word "phase" in this context and in equation (12.5). Consider the driver located at point $x = 0$. The closest element that is in phase with it (the first gray dot) has a phase shift of $\phi = -2\pi$ radians, since it is a full cycle behind the driver. The distance between these two elements is called

[3] The rate at which this energy is carried away is called the **impedance** of the medium.

[4] In fact, this is an example of a non-standard polarization and is important in the investigation of earthquakes. The two types of waves have different wave speeds and can be used to investigate the nature of the Earth's core (see here for details).

[5] A quick cheat for the speed of sound in air is $v = 331 + 0.6T_{\text{cel}}$ where T_{cel} is the temperature measured in Celsius. This can be derived from equation (18.4) and the binomial theorem (1.3).

the **wavelength** of the wave and represents where the wave profile repeats. The phase shift of all the other elements are given by the formula

$$\phi = -\frac{2\pi}{\lambda}x \tag{18.5}$$

where λ is the wavelength and x represents the distance from the driver. Using equation (12.5), we can represent the entire wave profile (in both space and time) with the following formula:[6]

$$\psi = A\cos\left(2\pi f t - \frac{2\pi}{\lambda}x\right) \tag{18.6}$$

The speed of the wave is directly related to the frequency and wavelength of the wave. We have

$$v = f\lambda \tag{18.7}$$

which follows directly from the constant velocity equation $x = vt$ and the definition of frequency $f = 1/T$.

Recognize that equation (18.6) only works for one-dimensional waves. For waves that propagate in multiple dimensions, the amplitude must decrease in order for energy to be conserved (energy is proportional to the square of the oscillation's amplitude). Some special mathematics is involved here . . . for a two-dimensional wave (like the ripples in a pond), equation (18.6) will involve the so-called Bessel functions. From the top the wave profile will be circular, but from the side it will have a shape (that oscillates up and down) like that in Figure 18.3.

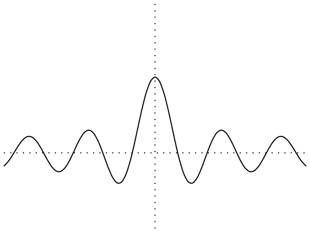


Figure 18.3: Bessel function of the first kind of order zero

In addition, we can talk about sources that are not simple oscillators: so-called **multipole** oscillators. The radiation patterns can be quite complex (see Figure 18.4 for some examples).

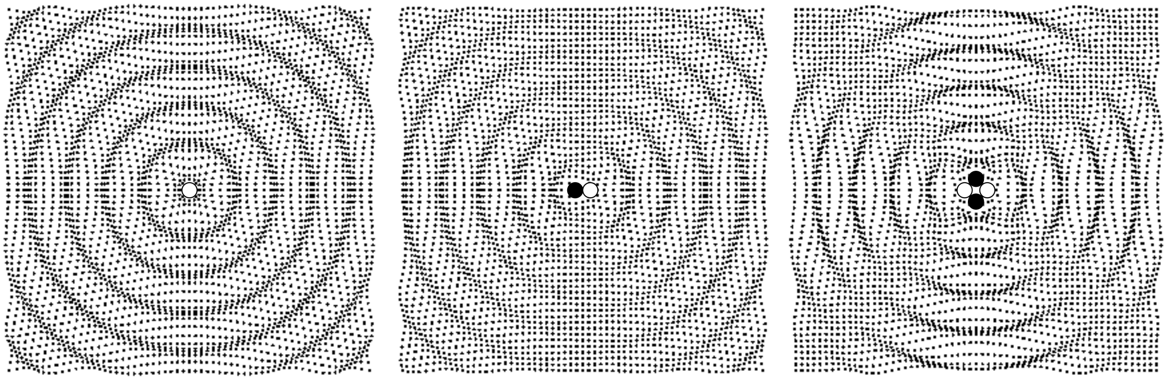


Figure 18.4: Radiation patterns from a monopole, dipole, and quadrupole sources (image credit here)

One interesting thing to note is that the radiation is directional: there are certain directions in which the radiation is minimal. This is one of the ways to construct directional antenna, for example. We won't need to know about this for our class however.

For the ideal string we have been considering, the formula for the wave speed is a constant. Frequently in other media this radiation speed will depend on the frequency of the disturbance. This is called **dispersion**.

One consequence of dispersion is that an arbitrary wave profile will not be preserved as it moves though the medium. In general, preservation will occur only if either (1) the medium is non-dispersive (not common), or (2) the wave profile is a pure sine wave—which has only a single frequency. What typically happens is that the sharp corners on the profile get rounded down over time and the profile spreads out due to its dispersion.

[6] This formula can also be written as

$$\psi = A\cos(\omega t - kx)$$

The quantity $k = 2\pi/\lambda$ is called the **wave number** and is the number of cycles per meter, similar to how frequency is related to period.

One final topic is required in this discussion of the radiation of wave energy: that of intensity. When we talk about sound and light as wave phenomena, the intensity corresponds to the our sensory perceptions of loudness and brightness (pitch and color correspond to the frequency of the wave). Physically, we define the **intensity** of a wave as the power that passes through a certain area divided by that area—a kind of density of how the power radiated.[7] For a single point source, the intensity is related to the distance from the source by

$$I = \frac{P}{4\pi r^2} \tag{18.8}$$

The intensity drops with an inverse square law because the surface area of the surrounding sphere increases with the square of the distance.

An important thing to note (for the next lecture) is that intensity is proportional to the square of the amplitude of the wave since the energy depends on the amplitude squared as in equation (12.9).

However, this definition of intensity does not quite complete the story. Both the sensation of sound and light are not linear—the ear and eye are more sensitive to lower intensities than higher intensities. This suggests creating an alternate scale to measure intensity called the **intensity level** and is measured in **decibels**. The formula for decibel level is

$$\beta = 10 \log(I/10^{-12}) \tag{18.9}$$

Although this scale could be used for light, it is designed to work for sound. This is because the sound intensity of 10^{-12} watts is the typical the **threshold of hearing**. This formula calibrates the decibel level at the threshold of hearing to zero. Every decibel increase corresponds to a ten-fold increase in power.

Frequently we need to convert from decibels to intensity. This is done with logarithms, but a short-cut involves using the inverse of the previous formula:

$$I = 10^{0.1\beta-12} \tag{18.10}$$

Next week we continue our discussion of wave motion. We investigate the mathematics of the interference of two waves and what it depends upon. We will also see that waves can rebound off of a boundary (reflection) and can interfere with the original wave source. For just the right combination, the incoming and reflected waves can combine to create a so-called standing wave. We will finish the topic by touching on the question of how waves of differing frequency interact.

[7] Actually, the intensity is closely related to the energy density of the field. The relationship is $I = ev$, where e is the density of energy (both potential and kinetic) and v is the speed of the radiation. We will see this again in Lecture 26.

Lecture 19

Wave Motion and Interference

Read sections 17.1–17.7

Having discussed the way in which waves propagate through a medium in the last lecture, we are now prepared to discuss how waves interact. Each wave emanates from a source and when the disturbances from two (or more) sources combine, the instantaneous amplitudes add.[1] In a formula one can say:

$$\psi = \psi_1 + \psi_2 \tag{19.1}$$

One important consequence of this is that if the two waves are sinusoidal (i.e., can be written in a form like equation (18.6)), the resulting disturbance will be another sine wave if and only if the original two have the same frequency. So for now we restrict our analysis to interfering waves with the same frequency. For simplicity we will also only consider waves with the same amplitude.

In general when two waves combine, the amplitude of the combination varies. This variation is called **wave interference** and is the quintessential characteristic of waves. When Thomas Young showed that light could be made to interfere with itself,[2] this eliminated the simple particle model of light which went back to Newton.

The interesting thing is that two waves can annihilate one another. This occurs when the two waves are out of phase with one another at the point of observation and is called **destructive interference**. When the two waves are in phase we get **constructive interference**. So, the amplitude of the resulting oscillation at the point of interference may be anywhere between zero and twice the original amplitude.

Remember that the energy of a vibration is proportional to the square of the amplitude according to equation (12.9). This implies that the energy transported to this spot may be anywhere between zero to four times the energy of the sources! In some cases (destructive interference) we have no energy transport, in other cases (constructive interference) we have twice the total source energy. Overall, the two extremes average out over the whole space yielding an energy balance.

So, the wave interference depends on the phase difference between the two vibrations. If $\Delta\phi = 0°$ we have constructive interference because the two vibrations are acting together. If $\Delta\phi = 180°$, we have destructive interference because the two vibrations are opposing one another. Any phase angle in between will cause something between these two extremes.[3] Technically it is possible for the phase angle to exceed 360° . In this case, the pattern simply repeats. Table 19.1 summarizes these conclusions.

[1] This is sometimes called the "principle of superposition". But, to me, this adds a lot of drama over something very simple. Besides, there are many superpositions principles (e.g., the electric field). All it means is that when two things combine they add. If the waves are polarized, this may involve a vector addition.

[2] You'll see how this is done in Lecture 21.

[3] For example, a phase shift of 120° will create an amplitude equal to the incoming amplitudes.

Table 19.1: Interference criteria, where n is any integer

$\Delta\phi/360°$	Interference Type
n	Constructive
$n+\frac{1}{2}$	Destructive

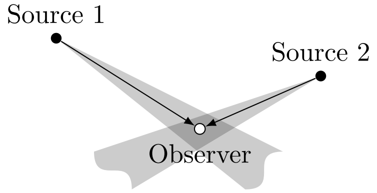


Figure 19.1: Two interfering wave sources

So far we have been discussing the interaction of waves at a certain point in space. Now consider two sources separated in space each sending a wave to a third point in space as in Figure 19.1. The interference the observer experiences is from the phase difference at the point of observation.

But we know that the phase of each wave will differ from their source according to equation (18.5). In addition, the two sources may not be in phase with one another. The interference type at the observation point is a combination of all these factors. The total phase difference is therefore

$$\Delta\phi = \frac{2\pi d_1}{\lambda} - \frac{2\pi d_2}{\lambda} + \phi_0$$

where d_1 and d_2 are the distances from the sources to the observer and ϕ_0 is the phase difference between the sources.

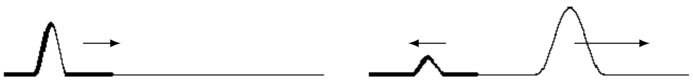
If we assume the initial phase difference is zero ($\phi_0 = 0$), then we will have constructive interference when $d_1 - d_2 = n\lambda$. This simply says that an integer number of wavelengths need to fit into the path-length difference between the sources and observer. The key is to calculate the number of wavelengths that fit into the path-length difference:

$$\frac{d_1 - d_2}{\lambda}$$

If this number is an integer, we have constructive interference (assuming the sources are in phase). If this number is a half-integer, we have destructive interference. In this way most of these interference type questions reduce to analyzing the geometry of the situation.

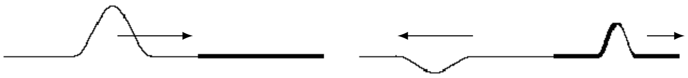
We now turn to a different topic: wave **reflection**. Whenever a wave encounters a sharp change in its medium it will generally split in two. One part will continue on into the new condition, but the other part will bounce off the interface and return. The first part is said to be the **transmitted wave**, the second is the **reflected wave**, and the initial incoming wave is the **incident wave**. If the medium increases in its wave speed, the amplitude of the transmitted wave will increase. This surplus is compensated by a reflection in the opposite direction. See Figure 19.2.[4]

Figure 19.2: Reflection in a string with 2x wave speed



If, however, the wave speed decreases, the transmitted amplitude will be smaller. The deficit is compensated in the reflected wave, which now has the opposite polarity. These two negatives (backward motion, opposite polarity) make a positive and balances the initial amplitude. See Figure 19.3.

Figure 19.3: Reflection in a string with 1/2 wave speed



So far we have been talking about intermediate cases. The extremes represent boundaries which may be either hard or soft. A **hard boundary** (or fixed end)

[4] The following two images are taken from here, which is a really great website to help understand vibrations and waves.

will reflect with the opposite polarity and represents the extreme version of Figure 19.3. A **soft boundary** (or open end) will reflect with the same polarity.

So far we have been talking about a pulse disturbance. If we now turn to discuss a wave like that represented in equation (18.6), the reflected wave can run back into the region of the incident wave and they may interfere with one another. In effect, the boundary acts as a second source.

In a one-dimensional situation between two sources (with the same amplitude and frequency) a special thing occurs. Since the wave energy is coming from both directions equally, a **standing wave** develops: the wave profile is fixed.[5] The standing wave continues to oscillate up and down, but the profile does not move at all. This standing wave occurs regardless of the distance or phase differences between the sources. The points of destructive interference are the **nodes** (these elements of the medium don't move at all), and the points of constructive interference of the standing wave (these are the elements with the largest movement) the **anti-nodes**.

Now consider a string of length L with one end fixed and at the other end a source of vibration (with frequency f). As the initial disturbance propagates through the string, it will reflect off the fixed end and form a standing wave. But when the reflected wave reaches the other end (with the original source) what happens? It will reflect again. But this time there is a difference—this "third" source is right on top of the first. Only if the geometry creates a situation in which these two constructively interfere will the standing waves accumulate. In other words, unless the situation is just right, the reflections will work against the vibration source.

But if the secondary reflection is constructive, the standing waves will resonate. And this is how every stringed musical instrument works. The required condition is simply that the wave repeat after it has traveled the length of the string and back. In other words,

$$n\lambda = 2L$$

We can rewrite this in terms of the vibration frequency using equation (18.7):

$$f_n = n\frac{v}{2L} \tag{19.2}$$

These are called the frequency **harmonics** of the string. The first harmonic, f_1 is called the **fundamental frequency**. Since the speed of propagation is related to tension via equation (18.1), we can see how tuning a guitar string can manipulate its resonant pitch and how different string weights can extend the guitar's range of notes.

Another one-dimensional class of musical instrument are the wind instruments. These set a column of air in motion and can be analyzed in a similar way. The main exception is that frequently these instruments are open at one end. This means that the reflected compression wave has the same polarity as the incident wave. But when the second reflection occurs the polarity flips (hard reflection).[6] We are off by half of a wavelength in order to get the standing waves to resonate. The formula for the harmonics of a half-open column of air can be written as

$$f_n = (2n-1)\frac{v}{4L} \tag{19.3}$$

A column of air open at both ends follows the same logic as the string, so it is also governed by equation (19.2).

The third major class of musical instruments is the percussion instruments, which involve banging things together. Although we could consider the triangle as a one-dimensional percussion instrument, usually we think of the drum. The vibration of a two-dimensional membrane is complex and very interesting.[7] Essentially, the

[5] Mathematically this is because the combined wave can be decomposed into a space and time factors. The wave profile is captured in the space component which does not move, and the oscillation is captured in the time component which does not depend on location. You'll need this to prove it.

[6] We were able to ignore this in the previous paragraph because the two hard reflections flipped the polarity twice.

[7] This website is a fantastic resource on the topic.

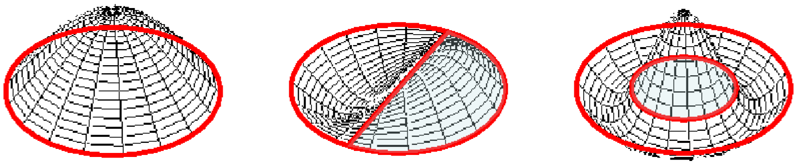


Figure 19.4: Standing waves on a circular membrane

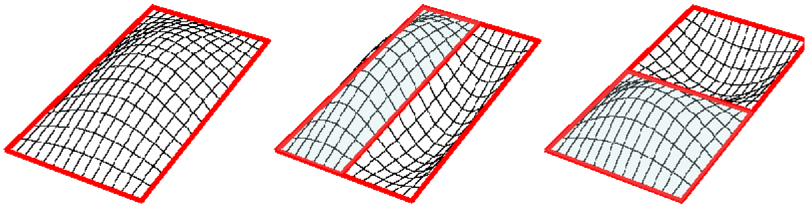


Figure 19.5: Standing waves on a rectangular membrane

geometry of the boundary will determine the way the waves are reflected back into the membrane.

The main constraint is that the membrane cannot move on this boundary (this is called a **nodal line**). The notion of a one-dimensional harmonic is replaced with the idea of a vibrational **mode**. Each mode has its own frequency and is primarily characterized by its nodal lines. Since these membranes are two-dimensional, each mode is characterized by two integers (for example, the number of circular and radial nodal lines).

We can talk about standing waves in three dimensions (applications to sound production and echoes are relevant here). In this case we need the mathematics of spherical harmonics which are classified by three integers.

It is even possible to consider the four quantum numbers used in defining the structure of the atomic orbitals as defined by standing matter waves (see Figure 19.6). There are four numbers in this case that correspond to the four dimensions of space-time. But we will leave that discussion for Lecture 27.

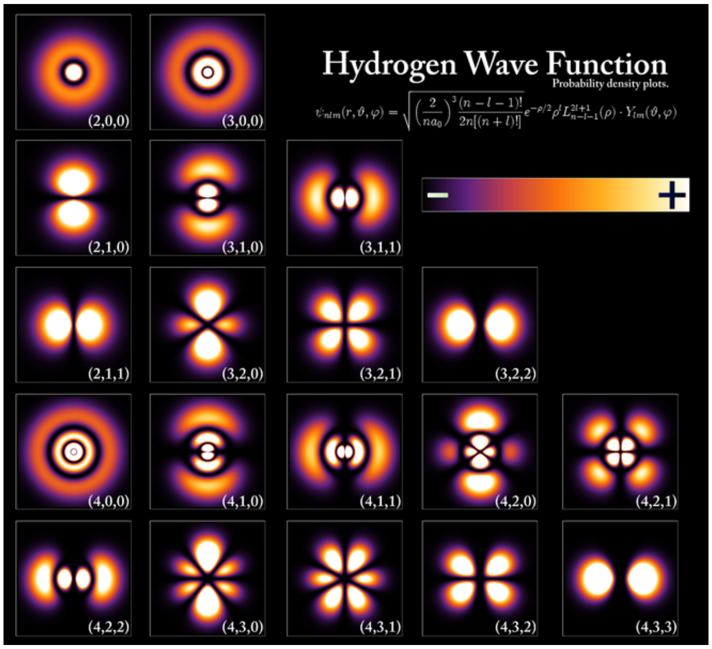

Hydrogen Wave Function
Probability density plots.
$\psi_{n\ell m}(r,\vartheta,\varphi) = \sqrt{\left(\frac{2}{na_0}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]}}\, e^{-\rho/2}\rho^l L^{2l+1}_{n-l-1}(\rho)\cdot Y_{lm}(\vartheta,\varphi)$
−
+
(2,0,0)
(3,0,0)
(2,1,0)
(3,1,0)
(3,1,1)
(2,1,1)
(3,2,0)
(3,2,1)
(3,2,2)
(4,0,0)
(4,1,0)
(4,1,1)
(4,2,0)
(4,2,1)
(4,2,2)
(4,3,0)
(4,3,1)
(4,3,2)
(4,3,3)

Figure 19.6: Atomic orbitals of the hydrogen atom (image credit here)

This nearly completes our initial survey into wave motion. There is one more topic worth discussing. At the beginning of this lecture we decided to focus on wave interference from sources with the same frequency. But equation (19.1) works whether the frequencies are the same or not. If we have two sources with different frequencies, they will generate **beats**. This is easiest to understand if the two frequencies are similar. Initially, the two sources are in phase and constructive interference occurs. Over time, however, the slower frequency begins to lag behind until the two are completely out of phase. This cycle of constructive and destructive interference continues. The amplitude of the interference cycles in time (see Figure 19.7). This **beat frequency** can be shown to be simply the difference between the two original frequencies.

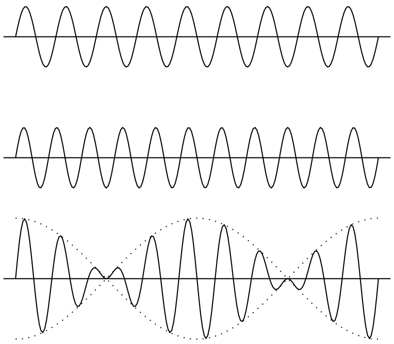


Figure 19.7: Two waves of similar frequency will combine as non-sinusoidal beats

We can extend this idea of beats to as many frequencies as we want. The resulting wave profile will be further and further removed from the simple sine and cosine functions with which we are familiar. However, we can turn this to our advantage using Fourier analysis.

We use **fourier analysis** to break an arbitrary profile into its sine wave components. Each component can then be analyzed separately. Combining them all back together yields the net motion for the wave profile under investigation.

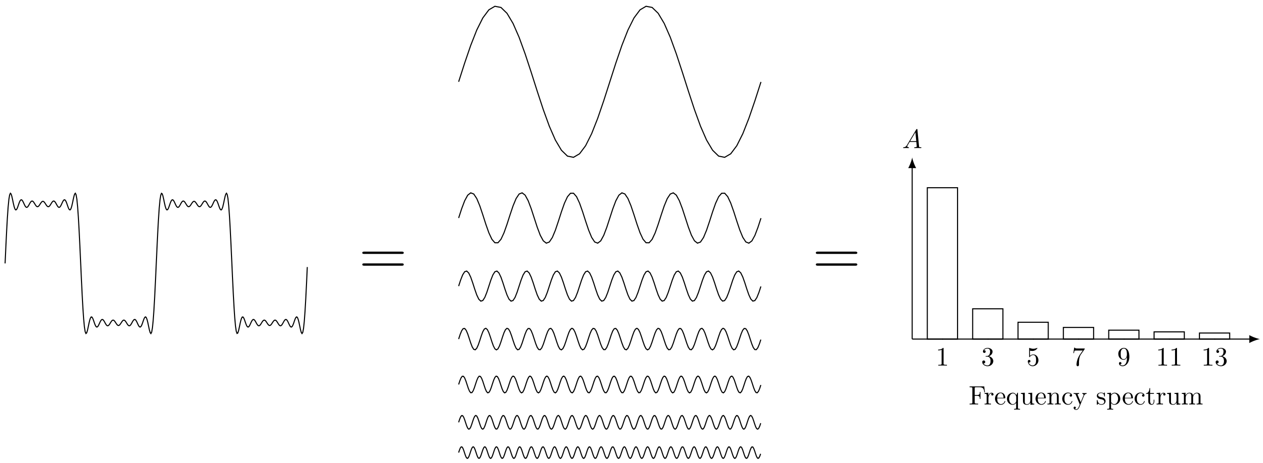


Figure 19.8: Example Fourier analysis for a step wave profile

Figure 19.8 shows how a step profile can be decomposed into sine wave components. The analysis determines the amplitude required for each component to combine into the desired shape. The sharpest corners correspond to the highest frequency components—which typically move fastest. This is why these corners are first to wash out in the profile. The highest frequencies determine how faithfully a digitized recording of sound can reproduce the original.

Next week we will move on to discuss light. We will examine the relationship between the wave motion we have been discussing and the ray picture of light. By doing this we are implicitly ignoring the wave nature of light (which is valid when the dealing with objects larger than the microscopic. This is called geometric optics and is governed by three simple laws regarding the propagation of rays, their reflection and refraction. We will investigate the operation of mirrors and lenses and how to correct their weaknesses. Finally we will quickly touch on Fermat's principle which is closely related to the principle of least action.

Lecture 20

Geometric Optics

Read sections 25.1–25.6 and 26.1–26.9, look back at 16.10

The study of light is one of those subjects that go back to antiquity. It's truly marvelous that a subject so common and essential to our daily lives would have such a deeply profound character. We now know that light is waves of electromagnetic energy governed by the laws of quantum electrodynamics. As such, light has a connection to all of the areas of modern physics: electromagnetism, relativity, and quantum mechanics.

For now, we will ignore these fundamental facts and focus on the way light moves. We will do this by using the idea of a **light ray**. In Lecture 18 we introduced the idea of waves as the radiation of energy. In this lecture we will make that idea more concrete.

When we draw the radiation of wave energy from a point source, it is common to do this with slowly increasing arcs (see Figure 20.1). Mathematically, each arc represents all the points of the wave pattern with the same phase. As time progresses, each point along this line oscillates up and down together. These lines are each called a **wave front** of the radiation pattern.

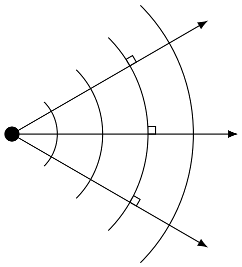


Figure 20.1: Wave fronts and rays from a point source

The light rays are defined as the lines that propagate through each wave front perpendicularly. Remember in three dimensions these wave fronts are actually surfaces of constant phase, so the rays are perpendicular to this surface.[1] One important consequence of this definition is that in variable media, the rays will curve (see Figure 20.2).

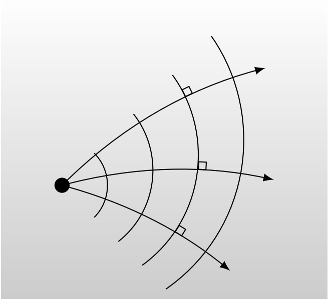


Figure 20.2: Wave fronts and rays from a point source in a variable medium

In variable media, the speed of the waves changes therefore so does the wavelength according to equation (18.7) (the frequency is constant). The wave fronts in slower media will be closer together which has a tendency to pull the rays into this area. This is called **refraction** and can occur for any multi-dimensional wave. This bending of wave energy occurs for earthquake waves as they move through the earth. This is one of the main ways we know that the earth's core has both liquid and solid layers (see here for more info). It also explains why mirages can sometimes be seen on a hot day and how fiber-optic cables work.

So far we have made no simplifying assumptions in our description. This approach using rays is mathematically equivalent to the use of the wave equations from the previous two lectures.[2] This explains why it was possible to consider a particle theory of light for so long, since it is easy to imagine particles of light moving along trajectories defined by these light rays.

Also, the wavelength of light is on the order of a micrometer or less. This means that the typical wave characteristics like diffraction and interference are only manifest when dealing with microscopic objects. In every-day applications we

[1] We can create a **propagation vector** which points in this direction and defines the ray completely if we give it a magnitude equal to the wave-number as defined on page 107.

[2] There is actually something profound in this which is related to the "wave-particle duality" we will see in Lecture 27.

can ignore the wave nature of light and assume that light rays obey the laws of **geometric optics**:

- Rectilinear propagation. In a uniform transparent medium without obstacles, light rays move in straight lines. (This is why shadows are sharp).
- Law of reflection. When encountering a reflective surface, light rays will bounce so that the angle of reflection equals the angle of incidence.
- Law of refraction. When encountering a change in transparent media, the refraction angle of the ray is governed by Snell's law (20.4).

We will discuss refraction shortly—for now we focus our discussion of reflection. In and of itself, reflection is pretty easy to understand: it's not unlike a ball bouncing off of a wall. Where things start to get interesting is when we use this reflection to focus light with the use of a **lens**.

Circular lenses are not the optimal cross-section for a lens, but circular lenses are much easier to make and they work reasonably well if the size is not too large.[3] This is called the "thin lens approximation".

If we take a small circular mirrored lens, it will focus incoming rays of light at a point halfway between the center of the circle and the far edge of the lens (see Figure 20.3). This is because the angle of reflection is measured off of the perpendicular from the mirror which goes through the center. This is called the **focal point** of the lens.

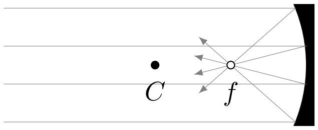


Figure 20.3: Focal point of a circular mirrored lens

The defining characteristic of the focal point is that parallel rays will meet at the focal point after being reflected off the mirror. We can use this fact to determine the location of the image reflected from the mirror.

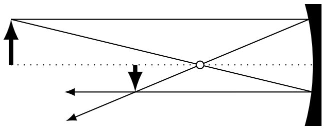


Figure 20.4: The image of an object outside the focal length is inverted and real

Consider Figure 20.4. The object is on the left and we consider two very particular light rays which leave the tip of the arrow. The first ray runs horizontal and we know that this ray will bounce through the focal point and continue on down. The second ray goes directly to and through the focal point. We know that this ray will bounce and then run horizontal because the geometry is the same whether the rays run backward or forward. These two lines intersect at the location of the image. If we were to put a screen here, we would get a sharp image of the object. The rays of light appear to be generated from this spot and it is therefore called a **real image**.

After examining the geometry of this situation (the upper and lower triangles are similar), the **lens equation** that governs this is

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f} \tag{20.1}$$

where d_o is the distance from the object to the lens and d_i is the distance from the image to the lens. The magnification of the height of the image is

$$m = -\frac{d_i}{d_o} \tag{20.2}$$

The negative sign indicates the image is inverted.

Though equation (20.1) is the commonly used lens equation, it can also be shown that

$$x_o x_i = f^2 \tag{20.3}$$

where x_o and x_i are the distances from the focal point to the object and image, respectively. This form of the lens equation is due to Newton (the other is associated with Gauss) and to my mind a bit easier to derive from the ray diagram in Figure 20.4.

[3] The best cross-section is a parabola which is why high precision instruments like radio telescopes are this shape. The headlights on your car are this shape too.

After playing with these equations for a while, you may notice that equation (20.1) gives a negative image distance if the object is within the focal length of the mirror (i.e., $d_o < f$). This is still correct and is telling us that the image is located on the other side of the mirror (see Figure 20.5).

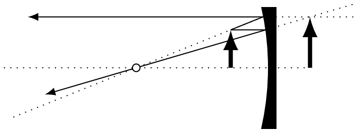


Figure 20.5: The image of an object inside the focal length is upright and virtual

In this case the image is said to be a **virtual image** because although the light rays appear to be coming from this point in space, they never do go there. Putting a screen behind the mirror will not capture a sharp image of the object no matter how hard you try. Notice that the magnification equation (20.2) still works because the negative image distance yields a positive magnification which implies the image is upright.

So far we have only discussed lenses that are convex. All the same principles apply for a mirror that is concave in shape (see Figure 20.6). For a concave mirror, the focal length is negative since parallel lines will diverge giving the impression that the image is behind the mirror. In fact, the image in a concave mirror is always virtual.

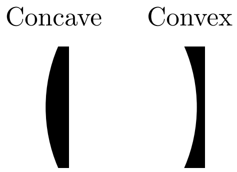


Figure 20.6: Lenses concave and convex

We can continue and talk about combinations of mirrored lenses, but the mathematics always involves using equations (20.1) and (20.2). The important point to remember is that the image of the first lens becomes the object of the next. In this way we can daisy-chain our way through any combination of lenses.

We are now ready to discuss the third law of geometric optics: refraction. Though the ancient Greeks knew about the law of reflection, they never could quite get the law of refraction. The correct law (known as **Snell's law**) was discovered in the early 1600s:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{20.4}$$

This governs the angles involved when a light ray encounters a sharp boundary in the transparent medium. The values of n_1 and n_2 are called the **index of refraction** for the material and the angles θ_1 and θ_2 are measured from the normal line to the surface (see Figure 20.7).

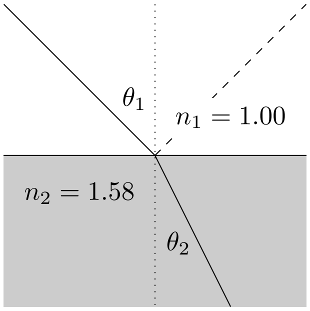


Figure 20.7: Snell's law of refraction

The index of refraction is a measure of the "refractive power" of the material. The higher the value of n, the more light will bend when entering it from the air. Though values can vary based on the composition of the material, typical values are listed in Table 20.1. By definition, the index of refraction of vacuum is one.

Air	1.0
Water	1.3
Glass	1.5
Diamond	2.4

Table 20.1: Typical indexes of refraction

If we shine a ray of light from inside a larger index of refraction material to a smaller one, the ray will bend outward according to Snell's law (20.4). There is a particular angle (called the **critical angle**) in which the refracted ray is effectively at 90° to the surface. Any incident angle beyond this is said to suffer total internal reflection because none of the light energy is transmitted out. The measurement of the critical angle for a substance is an easy way to accurately determine its index of refraction.

As we noted at the beginning of this lecture, the refraction of a wave is related to the propagation speed of the medium. This is true of the index of refraction as well. The relationship is simple:

$$n = c/v \tag{20.5}$$

where c is the speed of light in vacuum and v is the speed of light in the material.

Of course we can use the refractive power of a material to create a lens too. In just the same way as with a reflective lens we define the focal point to be the location where parallel lines meet after passing through the lens. Interestingly because this definition is the same as with mirrors, the geometry in the ray diagrams is very similar and the equations (20.1) and (20.2) also hold for thin transparent lenses also. The one difference is that the sign conventions are different because images from a lens tend to be on the opposite side of the object. Table 20.2 outlines the differences.

Table 20.2: Sign conventions for use in the lens equations

This is positive when...	Mirror	Lens
Focal length (f)	Convex	Convex
Object distance (d_o)	To the left	To the left
Image distance (d_i)	To the left	To the right

Just as with multiple mirrored lenses, we can combine transparent lenses and analyze the images using the lens equations. As before, the main principle is that the image of the first lens becomes the object of the next. Most microscopes and telescopes involve multiple lenses in order to magnify the objects under investigation.

One simple example is from optometry. The **optical power** of a lens is defined as the reciprocal of its focal length (so stronger lenses have higher numbers). This refractive power is measured in **diopters** which is simply one over one meter. When two lenses are touching, the refractive powers add.[4]. Since glasses and contact lenses are typically quite close to the lens of the eye, this simplifies the analysis of correcting vision by simply adding or subtracting diopters from the natural power of the patient's eye. The optical power of a normal eye is about 60 diopters.

Isaac Newton was very interested in the study of light, performed many experiments, and wrote a major treatise on the subject. And it was he who discovered the principle behind the rainbow: that white light is actually all the colors of the rainbow mixed. We can understand this realizing that most transparent materials are dispersive (that is, the speed of wave propagation varies with frequency). Since the color of light corresponds to the frequency of the wave, we expect the index of refraction to be color-dependent as well. Typically differences in color account for about a 2% change in the index of refraction.

We have already noted that circular lenses will only focus the light along a relatively small arc of the curve. If the arc is extended, the curve is too fast and pulls the rays in and away from the focus. This "de-focused" effect is called the **aberration** of the lens and is generally a thing to be avoided. The shape that will focus all the parallel rays of light perfectly is a parabola and has no aberration due to its shape.

However, another way lenses are aberrant is due to dispersion. Aberration of this sort is called "chromatic". It is sometimes possible to correct this kind of aberration by combining lens of different indexes of refraction to bring all the colors of light back together at the final focal point.

Other than equation (20.5), we have essentially ignored the speed of light. Nothing in the laws of geometric optics touches on the speed of light. Perhaps this is because light speed is so fast. For all intents and purposes it travels instantaneously. Of course we now know this is not the case, but geometric optics is just that: geometric. There is no time component involved, no dynamics.

In the late 1600s, Pierre de Fermat discovered that all three geometric laws could be summarized in one simple principle: the **principle of least time**. The time required to travel from point A to point B is along any other path than the true one takes more time. This was the first (and simplest) example of a variational principle in physics and became an archetype for several others including the principle of least action which we mentioned in Lecture 7.

Next week will bring our study of waves and light to a close. We will discuss the wave nature of light—including diffraction and interference phenomena. Nearly all of the concepts have been already introduced, but some interesting applications will be reviewed. We will mention how light is polarized. Also we will touch on

[4] This follows directly from equation (20.1).

the distinction between coherent and incoherent light—which is the difference between laser light and ordinary light. Finally we will finish the discussion on a philosophical note regarding the existence of the ether.

Lecture 21

Physical Optics

Read sections 27.1–27.7 and 27.9 then 24.1–24.2 and 24.6

In the previous lecture we focused on the properties of light that ignores its wave-like nature. This left us primarily with the simple radiation of energy which we found convenient to describe using the light ray. In this lecture we broaden this scope to discuss how light is like a wave.[1]

There are three observable properties that are associated with waves: interference, diffraction, and attenuation. Of the three really only interference requires a wave, but diffraction and attenuation are easier to understand with waves.

In the early 1800s, Thomas Young published his definitive experiment which provided evidence for the wave nature of light. We will start by describing his **double slit** experiment. (Quantum mechanics casts a shadow over the interpretation of these results—but we will leave that for the next lecture.)

The double slit experiment is simply that: two parallel slits close together. If we send a single coherent wave through these slits, they both act like waves sources that are in phase with one another and interfere. If we place a screen in the distance, we can capture the pattern of constructive and destructive interference.

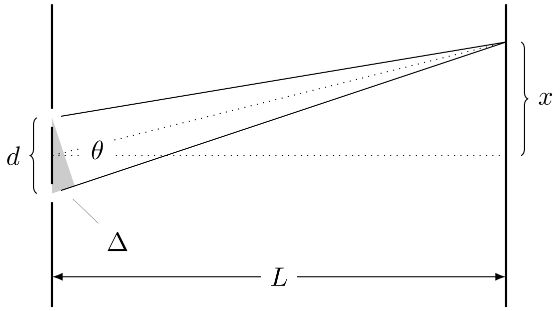


Figure 21.1: Geometry for the double slit experiment

Consider Figure 21.1. In particular the gray triangle which highlights the path-length difference, Δ, between the two rays. If L is sufficiently large in comparison to d, we can consider this to be right triangle with d as the hypotenuse. The angle opposite to the path-length difference is the same as the angle θ. This means that

$$\Delta = d \sin\theta$$

Based on the discussion in Lecture 19, we know we will have constructive interference whenever this distance is an integer number of wavelengths. Thus, we expect to see a bright band on the screen if the angle is given by

$$\sin\theta = n\lambda/d \tag{21.1}$$

[1] There is a third level of analysis awaiting us: the electromagnetic nature of light. We will touch on that in Lecture 26.

Of course, this same angle is given by $\tan\theta = x/L$. So we can reverse equation (21.1) to determine the wavelength of this light by measuring the distance x.

The main obstacle to overcome when observing the interference of light is the small wavelength. The range for the human eye is about 390 to 750 nanometers for blue and red light respectively. The interference pattern depends on this wavelength so it is typically very hard to see unless the geometry is of the same order. Also the wavelength dependence means that the pattern is best seen for monochromatic (single color) light.

The double slit experiment requires that light wave as the two slits be in phase. This is why firing two lasers at a screen will not produce interference. The way Young accomplished this was by placing a preparatory screen in front of the double slit. In this screen was a single slit essentially acting as a wave source for the double slit.

Another way to accomplish this is by taking splitting a coherent beam of light. This is usually done with a half-silvered mirror which reflects half the beam and transmits the other half. The arrangement in Figure 21.2 is called the Michelson **interferometer** is able to measure the speed of light with great accuracy. Essentially the interferometer is able to measure the distances between the two mirrors within 0.1 microns (the wavelength of light). Michelson and Morley originally designed this apparatus to detect the "ether wind" (a topic to which we will return in Lecture 26).

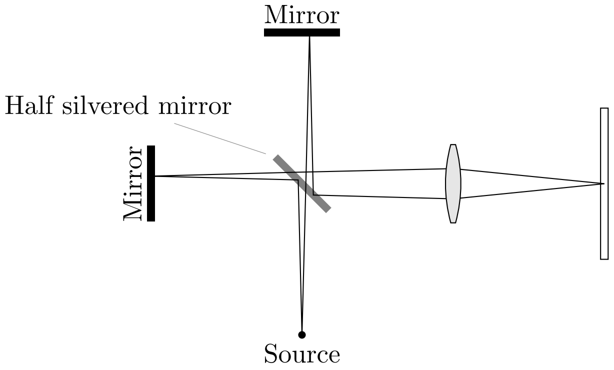


Figure 21.2: Schematic of the Michelson interferometer

Young's original double slit experiment depends upon the **diffraction** of light. This is the tendency for waves to bend around obstacles and why you can hear around corners. This is why the light bends and interferes in his double slit experiment rather than simply casting a very sharp shadow of the two slits.

If we were to remove a slit from Young's arrangement, the remaining slit will cast its shadow on the screen. But since the slit is small, the shadow will be fuzzy due the diffraction. In fact, the intensity of the diffraction pattern is given by[2]

$$I = \left[\left(\frac{\lambda}{\pi a\theta}\right)\sin\left(\frac{\pi a\theta}{\lambda}\right)\right]^2 \tag{21.2}$$

where a is the width of the slit (see Figure 21.3). This function assumes that the distance to the screen is much larger than the slit width so that the rays of light are effectively parallel from the slit.[3]

The diffraction given by equation (21.2) drops to zero whenever

$$\theta = n\lambda/a \tag{21.3}$$

except when $n = 0$ which is the brightest point of all. So even with a single slit we can see the wave nature of light. However, the side lobes of equation (21.2) are usually very faint, so effectively the image is spread over the range defined by (21.3) with $n = 1$.

[2] The function $\sin(x)/x$ is called the sinc function and happens to be the Bessel function shown in Figure 18.3. The connection here is merely coincidental as far as I know.

[3] Technically, $L \gg a^2/\lambda$. This approximation is called Fraunhofer diffraction. For screens very close Frensel diffraction corrects the calculation.

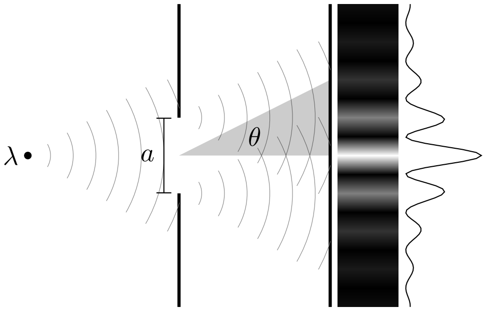


Figure 21.3: Single slit diffraction pattern (slit width exaggerated for clarity)

The previous analysis was all performed assuming the diffraction aperture was a thin slit. Another common type of aperture is a small circular hole. The diffraction pattern in this case is radial, but has the same characteristics as equation (21.2) except the pattern is slightly wider.[4] The width of the central diffraction peak is given by

$$\theta = 1.22\lambda/a \tag{21.4}$$

where a in this case is the diameter of the circular aperture.

Since the typical aperture in a telescope is circular, we expect the images of stars to exhibit just this kind of diffraction. An important consideration is the resolution of the images from two stars. We can use equation (21.4) as an objective measure of resolution. We say that we can distinguish two sources if the diffraction peaks do not overlap. This is called the **Rayleigh criterion** (see Figure 21.4). In fact, when using the human eye, the pupil is the effective aperture. Using equation (21.4) and 570 nanometers as the average wavelength of light, we can say that the human eye can resolve images with an angular separation of 1.4×10^{-5} radians, or about 100 arc-seconds.

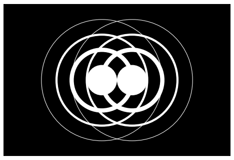


Figure 21.4: Point sources are resolved if the central diffraction peaks do not overlap

A **diffraction grating** is effectively a double slit analysis on steroids. A diffraction grating is good at displaying the wave nature of light because it separates the colors through interference rather than dispersion. Consider for a moment the geometry of the triple slit in Figure 21.5.

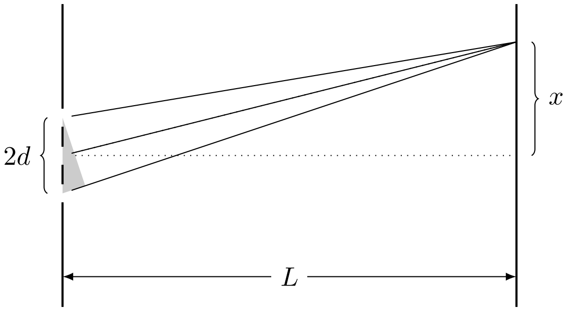


Figure 21.5: Geometry for a triple slit

The path-length difference emphasized by the gray triangle still governs the interference pattern. The lower slit introduces a third ray with an extra Δ of path-length difference. What this means is that when the double-slit angle produced constructive interference, the triple-slit also does. However, when the top two slits are in destructive interference, the third will propagate through. The diffraction pattern still peaks at the constructive interference points given by equation (21.1), but the peaks are much shaper (narrower and taller).

A typical diffraction grating will have much more than three slits. Typically we are told the number of slits per centimeter or some similar metric. The angle of diffraction is proportional to this number.

A diffraction grating will disperse light into a rainbow because equation (21.1) is wavelength dependent. A standard DVD produces this effect because of the closely spaced tracks on the disk. The regular spacing in crystals will also act like

[4] This radial pattern is called the Airy disk. Calculating the diffraction pattern from any aperture involves Fourier transforms which are closely related to the Fourier analysis we mentioned in Lecture 19.

a diffraction grating for certain wavelengths of light. This kind of analysis is how the double helix structure of DNA was first discovered.

Usually the iridescence seen in nature is not from a diffraction grating, however. It is much more common for this effect to be produced by **thin-film interference**. Simple examples include soap bubbles and the colorful sheen of oil on water.

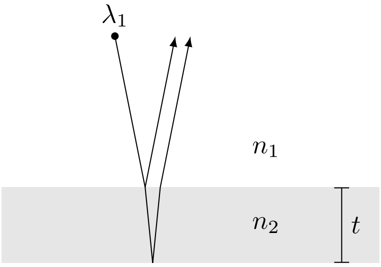


Figure 21.6: Thin film interference

Physically this occurs because both the inner and outer boundary act as reflection points. The light from the front of the layer will constructively interfere with the light from the back of the layer if the distances are just right (see Figure 21.6).

If the thickness of the film is t, the path-length difference is $2t$, so you might think we require $2t = m\lambda$ (we use m to avoid confusion with the index of refraction). However, the reflection from the front is a hard reflection because the refractive index of the film is larger than air. The reflection off the back is a soft reflection for similar reasons, so the interference acts something like the open tube from Lecture 19. This means we require

$$2t = (m + \tfrac{1}{2})\lambda_2$$

where λ_2 is the wavelength of the light in the medium. But we know that the wavelength of the light changes with any change in the index of refraction because its speed will change. We combine equations (20.5) and (18.7) with the previous condition to yield the thin-film interference formula:

$$2t = (m + \tfrac{1}{2})\left(\frac{n_1}{n_2}\right)\lambda_1 \tag{21.5}$$

where we give the wavelength the subscript 1 to emphasize that this is the source wavelength.

You might wonder why the film must be thin. The reason is that the two rays must be of comparable amplitude in order to interfere. Typically as light moves through a transparent medium, it loses intensity. This is called **attenuation** and is a kind of optical friction or resistance. It is possible to incorporate this attenuation by allowing the index of refraction to be a complex number, though we will not do so here. The attenuation of light causes the intensity to fall exponentially.

There is another characteristic of light that is a consequence of its wave nature: **coherence**. This is the difference between incandescent light and laser light. In laser light, each pulse of light is aligned by both phase and polarization.[5] This means that effectively all the light rays constructively interfere all the time. With an incandescent light, the alignment of phase and polarization is random: sometimes its constructive, sometimes its destructive.

Since the intensity of the light is proportional to the square of the amplitude, when the intensity of the incandescent bulb increases two-fold, the intensity of laser light increases four-fold. So, the high intensity of laser light is due to the coherence of its light—which is a wave phenomena.

Another piece of evidence for the wave nature of light is its polarization. The polarization of light was first discovered through **birefringence**: a material in which the index of refraction depends upon the polarization of the light. A typical image seen through this material will double as the light is polarized relative to its crystal structure.

The best way to explain why the polarization of light cannot be explained by particles is by considering two crossed polarizing filters. When we pass light through a filter, the component of the light wave aligned with the polarizer axis passes through. If we place a second polarizer at 90° to the original, it will block the light because there is no component left to pass through.

If, however, we insert a third polarizer in between them aligned at an angle, light will begin to pass through again. When two polarizers are at an angle θ, the

[5] This occurs because a type of standing wave is built in the lasing material which excites more coherent light via a quantum mechanical process. Although the production is based on quantum mechanics, the coherence of the light is understandable with classical ideas.

intensity of light that passes through is given by **Malus' law**:

$$I = I_0 \sin^2 \theta \tag{21.6}$$

This is because the intensity is proportional to the square of the amplitude and the polarizer allows a component $A = A_0 \sin \theta$ through.

If the middle polarizer is set at 45° then it will let half of the intensity of light through. The final polarizer is also at 45° to the middle, so one-quarter of the intensity of light is allowed through even though the two outer polarizers are crossed.

So, the evidence is pretty clear that light is a wave phenomena. Of course every wave does so in a medium and it is natural to ask what is the medium for light waves? In the 19th century this medium was given the name the **ether** and one of the main goals was to determine its mechanical properties.

But what is it? Like an elastic solid transmitting sound waves? This substance must be present everywhere light is—including the depths of outer space. Apparently it is everywhere, yet we do not experience any "ether drag" like we do with air. If we model this as an elastic solid, its wave speed is given by equation (18.2). But the speed of light is enormous. This implies a very rigid solid which is less dense than air—yet its presence is undetectable mechanically. The search for an "ether wind" was a hot topic around the turn of the 20th century.

As the evidence piled up, it began to be more and more difficult to imagine any mechanical model that could act as a medium for light. During this time, Einstein's special theory of relativity altered our views of space and time effectively declaring the ether concept dead. Now, with the advent of quantum field theory, we no longer even think of light as waves—but that's the subject for next week.

Next week we will continue discussing mechanical models for the ether. We will find that the more we refine this model, the more problems we uncover. This is one of three major breeches in classical mechanics. In this case, the problem is finally resolved in the theory of relativity. On the other hand, we have quantum mechanics as the resolution of the other two problems (kinetic theory and atomic theory). In this way we will finish the term recognizing the limits of classical mechanics.

Lecture 22

Limits of Classical Mechanics

Review Lectures 16 and 21. Sneak a peak at sections 28.6–28.7, 29.1–29.3 and 30.3

In the last lecture we left open the question on the ultimate nature of light. The evidence for some sort of wave picture is clear. However, the mechanical nature of the medium doing the waving was less clear. By the middle of the 19th century, some connection between electricity, magnetism, and light was understood. As electromagnetic theory matured (culminating in Maxwell's equations), the problem became worse.

One clear example is that light (as an electromagnetic wave) should carry momentum. The problem is that mechanical waves do not carry momentum. The transfer of momentum implies a transfer of mass: movement of the medium itself.

Well, perhaps the ether is a fluid. A fluid can easily carry momentum and support wave motion. In fact, a thought at the time held that elementary particles might be modeled as a vortexes in the fluid. Like a smoke ring, a vortex can maintain its structure over an extended period of time.

Unfortunately, a fluid cannot support a polarized wave. Because a fluid only has pressure stress (no shear or normal stress), the compression wave has only one degree of freedom. But electromagnetic theory and experimental results require a polarized wave theory of light.

Whether solid or fluid, we expect a basic result called the "ether wind". When you are riding a car and roll down the window, can you tell you are moving? Of course you can: the wind gives it away. The air is not moving: the wind is caused by your motion through the medium. In the same way, we ought to be able to detect our motion through the ether.

Of course, our motion needs to be comparable to the speed of light—which is fast. One trick was to measure this ether wind by using the motion of the earth. Around the turn of the 20th century, Michelson and Morley performed a number of experiments to detect the motion of the earth through the light ether. The now famous "null result" caused quite a bit of controversy at the time. Theory after theory was proposed[1] but the whole thing reminds one of the epicycles of Ptolemaic astronomy. Tweak after tweak attempting to fit a dying theory to unyielding facts.

The answer was summarized in 1905 by Einstein in the special theory of relativity. By the statement that $E = mc^2$, we can see that every wave moves mass through the energy it moves. For everything except a light wave, the amount of mass is

[1] For example, maybe the earth moves in some sort of ether bubble.

completely negligible. But for light, the momentum transfered follows from this formula.

However any mechanical model of light based on Newton's laws was in contradiction with the principle of relativity (see page 27). Because this medium is universal and everywhere, it acts as an absolute reference frame. But Einstein showed that it is possible to have our cake (electromagnetic theory) and eat it too (principle of relativity) and declared the ether dead.

The trade-off is that we must reassess our assumptions about space and time. Because of this the speed of light acts as a universal speed limit which implies that speed and velocity works in a way that is non-Newtonian. This explains the null result of Michelson and Morley, but at the expense of bizarre consequences like length contraction and time dilation. We will return to these ideas in Lecture 26.

Relativistic mechanics wasn't the only correction required to classical mechanics. In many ways quantum mechanics was an even more radical correction than relativity. Interestingly, Einstein played a large role in development of quantum mechanics too.

Looking back now, the first hint how wrong we were to assume the subatomic world operated under classical principles of mechanics was the **ultraviolet catastrophe**. This comes from the analysis of the spectrum of radiation from a heated object. The net power radiated is given by equation (15.3), but what we seek is the distribution of this energy across the electromagnetic spectrum. Classical theory predicts an ever-increasing amount of energy in the lowest wavelengths, which is nonsense.

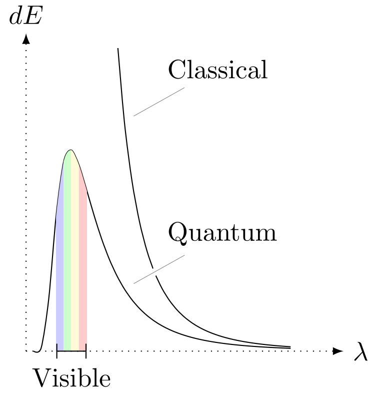


Figure 22.1: Ultraviolet catastrophe: Expected spectrum of energy by wavelength from a hot object. Both curves assume a temperature of 5000 kelvin—the approximate temperature of the sun.

In 1893, a result known as **Wien's Law** was empirically discovered:

$$\lambda_{\max} = \frac{2.898 \times 10^{-3}}{T} \tag{22.1}$$

In this formula, the peak wavelength in the spectrum is inversely proportional to the absolute temperature. Ultimately, quantum theory predicts a spectrum in line with reality (see Figure 22.1) and justifies Wien's law.

In both classical and quantum theory, the number of ways a medium can vibrate (i.e., its modes, see page 112) is proportional to the square of the frequency.[2] If the medium is in thermal equilibrium at a particular temperature, the equipartition principle from kinetic theory (page 96) implies that each mode will hold kT units of energy.

If the ether is a classical wave medium, this implies that the amount of radiant energy from a hot object should be dominated by the high frequency, low wavelength (ultraviolet) range of the electromagnetic spectrum—in other words, hot objects should be blue rather than red. Classic theory is not even close. Relativity doesn't save us here because the energies involved are no where near relativistic.

In 1900, Max Planck was able to mathematically reverse engineer the actual radiation spectrum and came to the conclusion that the problem was the equipartition principle. He found that with a single simple assumption, he could derive the correct spectrum:

$$\Delta E = hf \tag{22.2}$$

In other words, the modes are only allowed to exchange energy in increments that are proportional to the frequency of that mode. This has the effect of suppressing the higher frequency modes because the probability of a large energy exchange is exponentially small in thermal equilibrium.

Another phenomenon points to an unusual connection between energy and frequency[3] of light called the **photoelectric effect** first discovered in 1839. If one shines intense light on metal it will ionize and the electrons thus stripped off can

[2] This assumes the medium is three dimensional. For a string (one-dimensional), the number of modes is constant across frequency. For a sheet, the number of modes increases linearly with frequency.

[3] Remember: classically, energy in a wave is related to its amplitude not its frequency.

be captured in an electric current and measured. The odd thing is that the speed of the escaping electrons depends on the frequency of the light. In fact, for very low frequency light no electrons are released no matter how intense the light.

But why? A few years later, Einstein realized that both Planck's assumption and the photoelectric effect can be explained through a particle theory of light. Each particle, called a **photon**, has energy related to its frequency according to (22.2). It is actually this work rather than relativity for which Einstein received his Nobel Prize in 1921.

Einstein realized that the explanation of the photoelectric effect is simple if light comes in streams of photons. Each photon only carries hf worth of energy, so if the frequency is not high enough, the light particle does not have enough energy to lift the electrons out of the potential well that binds them in the atom. Increasing the intensity of light only increases the number of photons—none of which are powerful enough to move the electrons.

The most important thing about the photoelectric effect is that it cannot be explained with waves. This frequency dependence of energy is without parallel in the classical theory. The fact that energy transfer is quantized by (22.2) is unprecedented.

And there were other problems—particularly at low temperatures. The specific heat of solids predicted by classical theory was wrong (Einstein also proposed the first approximation of a solution using quantum ideas). Superconductivity was discovered in which the electrical resistance of a metal disappears. Liquid helium below about 2.17 degrees kelvin completely loses all viscosity (now known as a superfluid).

Beside issues with classical kinetic theory, deep inconsistencies were discovered in the structure of the atom. Classical electromagnetic theory implies that an orbiting electron will continually radiate energy. Based on the size of the atom, the theory predicts that all the orbital energy should be radiated away in 10^{-11} seconds or so. Atoms are a bit longer lived than that: another blatant contradiction between theory and reality.

Spectroscopy is the study of the spectrum of radiation from materials. It was discovered that the elements only emit radiation along certain frequencies: creating a pattern of lines rather than a smooth spectrum. Both this and the non-radiating orbital electron seem to imply that energy transfer is limited in the microscopic realm. The **Bohr model** of the atom (see Lecture 28) was successful in collecting these facts, but left unexplained the fundamental reason of why energy is quantized.

Out of this cacophony of confusion quantum mechanics was born. Nearly 30 years of struggle and debate finally gelled into a single explanation—which is the subject of Lecture 27. Einstein's photon idea combining particle and wave (in some obscure way) led Louis de Broglie to postulate that perhaps elementary particles (in particular: electrons) have both particle and wave characteristics too. These "matter waves" are critical in understanding the structure of the atom and form the foundation of chemistry.

So around 1925 order was brought back to the land of physics. Except one thing. Quantum mechanics is not relativistic. And the most straightforward way to combine the two was fraught with problems (like negative kinetic energy, infinite numbers of particles, etc.). It took another quarter of a century to realize that the conflict between quantum mechanics and special relativity was only on the surface. The development of quantum electrodynamics (QED) has been described as the "crown jewel of physics" by Feynman due to its astonishing accuracy (better than ten parts per billion so far).

The mathematical framework for QED is called quantum field theory (see Lecture

27). One important consequence of quantum field theory is the division of the world of elementary particles into **fermions** and **bosons**. The electron is an example of a fermion which ultimately explains the periodic table. The photon is an example of a boson which ultimately explains Planck's solution to the ultraviolet catastrophe. Fermions are the building blocks of matter and bosons are the carriers of force.

Quantum field theory has been utilized to explain the two nuclear forces as well. Though more complicated, these theories are also accurate to within current experimental precision. So the fundamental framework of subatomic physics is able to explain every experiment to date. This is called the **Standard Model** of physics (see Lecture 30).

But the holy grail, the "theory of everything", still eludes us. Oddly enough, it is gravity—the first force of them all—that stands alone. It took Einstein a dozen years to discover a relativistic theory of gravity (general relativity). No one has discovered a quantum theory of gravity. Stay tuned.

Lecture 23

The Electric Field and Potential

Read sections 18.1–18.9 and 19.1–19.5

Electricity, magnetism, and gravity are all long-range forces. Rather than acting via the direct contact of objects (like elasticity or friction), these forces act "at a distance". Though gravity was the first to be quantified in Newton's inverse square law (6.3), magnetism was the first to be studied. Perhaps due to its portability: one can play with magnets much easier than planets.

In 1783, it was discovered that the electric force also obeys the inverse square law[1]

$$F = \frac{kqQ}{r^2} \tag{23.1}$$

which is called **Coulomb's law**. Each q represents the electric charge on the object and k has a value of 8.9876×10^9 in SI units.[2]

Unlike gravity (and like magnetism), the electric force is both attractive and repulsive. This implies that there are two different kinds of electric charge. The convention to call them "positive" and "negative" goes back to Benjamin Franklin. One advantage to this convention is that equation (23.1) now tells us when the electric force is attractive (charges of the same charge) or repulsive (charges of opposite charge).

It is not hyperbole to say that our modern standard of living depends on the correct understanding and harnessing of electric power. In general, materials fall into two categories: **conductors** which easily allow the flow of electric charge and **insulators** which essentially block the flow.[3]

Take an object with some excess charge on it. If we touch this object with a conductor, some of the excess charge will flow into and through the conductor because all of the individual charges are repelled from one another. In general, any excess charge will distribute itself through a conductor in order to maximize the distance between the charges. This is also why any excess charge will reside on the surface of a conductor.

There is another, trickier, way to transport electric charge. Take the same object as before with the same excess charge on it. Move the conductor close to the object without touching it (or place a thin insulator between them). The charge in the object will repel similar charge in the conductor and attract the opposite charge. In this way, a separation of charge is induced in the conductor. If we drain the repelled charge (by connecting it to the ground—literally), we are left with a net charge in the conductor of the opposite polarity.

[1] All three long-range forces roughly obey an inverse square law. Interestingly, only the electric force obeys the inverse square law exactly.

[2] We will see why in Lecture 26, but this happens to be the speed of light squared times 10^{-7}.

[3] These terms are used in the context of anything that flows: for example, heat.

These tricks were essential to the initial study of electricity. They also reveal an important fact of nature: matter is electrically active. Of the three long-range forces, electricity is the most powerful. Ironically, this strength acts to conceal the electrical nature of matter. Negative electrons[4] are strongly attracted to the positive nucleus and lock into an overall neutral atom. Except for unusual (and dramatic) circumstances, electricity is only manifest on the microscopic scale.

But, indirectly, electricity is everywhere because it is the source of all elastic, contact, and chemical forces. In fact, every macroscopic interaction except gravity is ultimately electric. In most cases, we need quantum mechanics to fully understand how this is true (in particular, the solidity of matter and the covalent bond). The key insight is to recognize that any asymmetry in charge distribution can induce a similar asymmetry in adjacent atoms. This charge separation is idealized in the **dipole moment** which is the product of magnitude of the charges and the separation between them (see Figure 23.1).

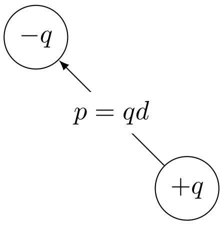


Figure 23.1: Dipole moment calculation

In the mid-19th century, Micheal Faraday introduced the idea of the electric field into the physics community. He found it necessary to conceptualize what was happening in certain electromagnetic experiments (we will see these in Lecture 25). A couple of decades later, Maxwell summarized the dynamics of the electromagnetic field in his four famous laws. Forty years later, Einstein realized that these laws are not compatible with Newton's laws of motion and developed the special theory of relativity. He later emphasized that the field concept was the key to uncovering the incorrect assumptions underneath Newton's laws of motion.

The field also occupies a middle ground for those uncomfortable with "action-at-a-distance" by offering an intermediate entity to carry the influence of the force. We take Coulomb's law (23.1) and divide it into two pieces

$$F = (q)\left[\frac{kQ}{r^2}\right]$$

which isolates (in square brackets) the part that is not associated with the charge experiencing the force. We call this the **electric field** generated by the charge Q. In short, the electric field from a point charge is

$$E = \frac{kQ}{r^2} \tag{23.2}$$

and we can rewrite equation (23.1) as $F = qE$ which makes clear why the SI unit for the electric field is newtons per coulomb. The field from multiple charges is simply the sum of the fields from each. Though recognize that the field is a vector (because force is) so some vector addition may be required.

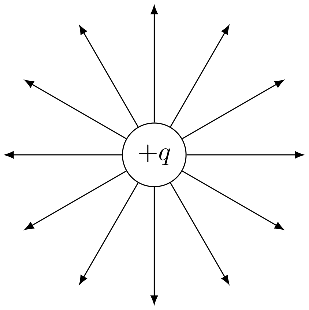


Figure 23.2: The electric field around a point charge represented by lines of flux

We can draw the electric field around a charge as in Figure 23.2. Each line represents electric **flux** which points in the direction of the force and the field strength is represented by the density of flux lines in the diagram. Lines of flux always point from positive charges and end with negative charges (or drift off to infinity).

The superposition of multiple charges can easily create quite complicated field line patterns. The simplest is produced by an infinite sheet of charge. Assuming the charge density is constant (which it should be if the charges are at rest), then by symmetry we can see that the field lines will simply be straight lines perpendicular to the sheet. Since the field lines don't diverge, this represents a constant electric field.

If we place a dipole in the presence of a constant electric field, the positive charge will be pulled along the field and the negative charge will be pulled against the field. In Lecture 10 we called this a force couple and the effect is to produce a net torque on the dipole. The total torque acting on the dipole is given by

$$\tau = pE\sin\theta \tag{23.3}$$

[4] There are about 6.25 trillion electrons in a coulomb, the SI unit of electric charge.

where θ is the angle between the dipole moment and the electric field. So the dipole will twist until it lines up with the field.

Suppose we place a second sheet of opposite charge parallel to the first. This creates a **capacitor** (see Figure 23.3). A capacitor is said to store charge since the charges on the two plates are held in place by their mutual attraction to one another. For a perfect (infinitely long) capacitor, there is no electric field on the outside. True capacitors always have **edge effects**, so not all the electric field is contained inside the system. Also, the two plates of a capacitor are typically separated by an insulator which allows a small amount of charge to flow and will eventually neutralize the charge separation. This is called **leakage** in the capacitor.

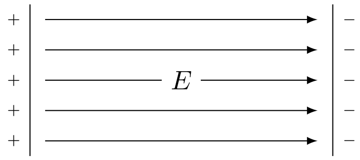


Figure 23.3: The electric field inside an ideal capacitor is constant

From Lecture 18 we know that the power intensity (18.8) from a point source falls off according to an inverse square law. The total energy is conserved but as the radiation propagates through space it is spread across a larger and larger surface area according to $4\pi r^2$. The mathematics is similar for the electric field from a point charge. This means that for the electric field, the total flux is conserved. This is summarized in **Gauss' law**:

$$\sum_{\text{surf}} \Phi_{\text{ele}} = Q_{\text{tot}}/\epsilon_0 \tag{23.4}$$

where Φ_{ele} is the total electric flux that passes through the surface of interest and Q_{tot} is the net charge inside the surface. The symbol ϵ_0 is called the **electric constant**[5] and is equal to $1/4\pi k$ which is 8.854×10^{-12} in SI units. One way to interpret this equation is that every line of flux starts with a charge—no lines of flux start (or end) in thin air.

We need to be careful to account for only the component of the field line that passes perpendicularly through the surface, so the formula for the flux is

$$\Phi_{\text{ele}} = (E\cos\theta)(A) \tag{23.5}$$

where θ is the angle between the field vector and the normal vector to the surface area A.

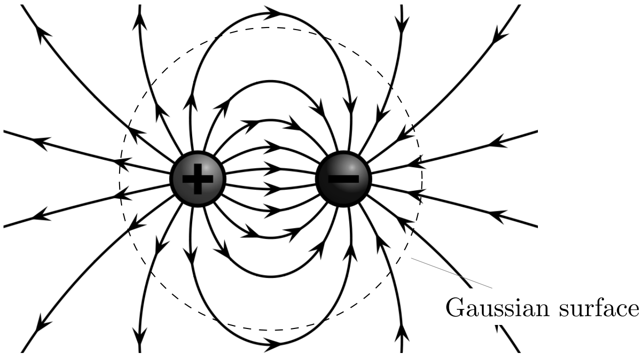


Figure 23.4: Applications of Gauss' Law to the electric field around a dipole (image credit here)

As an example of Gauss' law in action, consider the dashed circle in Figure 23.4. Notice how each field line that goes out is compensated by a line that goes in. The net flow of flux is zero which corresponds to the total enclosed charge being equal to zero.

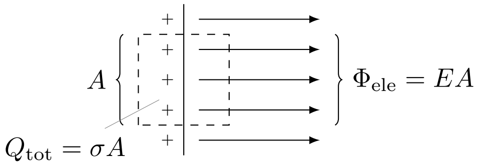


Figure 23.5: Calculating the electric field in a capacitor with Gauss' law

A more substantial application of Gauss' law is to calculate the magnitude of the field in an ideal capacitor. According to Figure 23.5, the total charge[6] enclosed

[5] I learned this under the name "permittivity of free space", but the "electric constant" is much easier to remember!

[6] Remember, all the charge is collected on the surface of the conducting plate.

in a rectangular box parallel to the positive plate is σA, where σ represents the charge density (i.e., $\sigma = Q/A$). The only flux contribution is from the face that is parallel to the surface, so the total flux coming out of the box is EA. Using Gauss' law (23.4) we can solve for the electric field as

$$E = \sigma/\epsilon_0 \tag{23.6}$$

Mathematically, we say that Gauss' law tells about the **divergence** of the field. Field lines are only created at positive charges (known as "sources") and they are consumed at negative charges (known as "sinks"). The field is also characterized by its **curl** which tells how much the field twists and turns as we follow the flux (think about a whirlpool in water). The way we evaluate this curl is by following a closed path and calculating the component of the field parallel to the path. The total accumulation of field along this path is the amount of curl in the field. The electric field has no curl because it comes from a central force. In symbols we have:

$$\sum_{\text{loop}} E_t \Delta s = 0 \tag{23.7}$$

where E_t is the component of the field parallel to the little bit of loop Δs we are at in the sum.

This should remind you of the discussion of conservative and non-conservative forces in Lecture 7 (see page 40). The electric force is conservative and a potential energy function exists. So far, we have drawn an analogy between intensity and flux. There is also a correspondence to this potential energy: the **field potential**. Both stem from the connection between field and force in the equation $F = qE$. The field potential is defined[7] as

$$\Delta V = -E_t \Delta s \tag{23.8}$$

In other words, the electric potential V increases as we move against the flux lines in the field. This is similar to pushing up a ball against the force of gravity. This formula also implies that the potential does not change as we move directly across the lines of force. The formula for the electric potential energy is simply[8]

$$PE_{\text{ele}} = -qV \tag{23.9}$$

The SI unit for electric potential is the volt. A common unit of energy at the microscopic level is the "electron-volt" which is the charge of the electron (1.602×10^{-19} coulombs) times a volt. This is also why the electric potential is also called **voltage**.

An **equipotential** surface is the collection of all the points in the electric field with the same value of potential or voltage. The previous paragraph implies that at each point the equipotential is perpendicular to the field lines. These equipotential surfaces are similar to the contour lines on a topographical map and the way wave-fronts and rays are related (see Lecture 20).

When we place a conductor in the middle of an electric field, the surface of the conductor will form an equipotential. The reason is that any voltage difference corresponds to a net electric force along the surface. But the charge in the conductor will flow until this pressure is equalized. This also implies that the field lines that touch a conductor will be perpendicular to the surface. The conductor pulls the field lines toward itself by redistributing its internal charge (see Figure 23.6) until they are perpendicular.

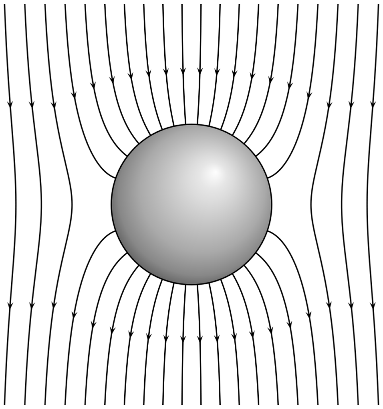


Figure 23.6: Adding a conducting sphere to a constant electric field will cause the field lines to bend into the conductor (image credit here)

The formula for the electric potential from a single point charge is

$$V = \frac{kQ}{r} \tag{23.10}$$

[7] Unfortunately, the negative sign here makes the signs in some calculations confusing—in particular the next equation (23.9). This sign convention propagates through all of electromagnetism, so beware of signs!

[8] Compare this with the gravitational potential energy (7.3): mgh. Apparently the gravitational field is simply $-g$!

From this formula, you can see that the equipotential surfaces around the charge are spherical shells of increasing radius.

The advantage of using the field potential rather than the field directly is that it is not a vector. So the total potential is the simple sum of the potentials from each charge. For example, the potential from a dipole is

$$V = \frac{kq}{r_+} - \frac{kq}{r_-} \tag{23.11}$$

where r_+ and r_- are the distances to the positive and negative charges respectively (see Figure 23.7).

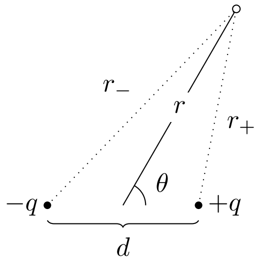


Figure 23.7: Calculating the electric potential of a dipole

If we are far from the dipole, these distances are nearly parallel and can be approximated by[9]

$$r_\pm = r \pm \tfrac{1}{2} d\cos\theta$$

where r is the distance from the center of the dipole and d is the distance between the charges. In addition, since r is much larger than d when we are far away, we can use the binomial theorem (1.3) to say

$$\frac{1}{r_\pm} = \frac{1}{r}\left[1 \pm \frac{\frac{1}{2}d\cos\theta}{r}\right]^{-1} = \frac{1}{r}\left[1 \mp \frac{\frac{1}{2}d\cos\theta}{r}\right]$$

Putting this result into equation (23.11) gives us the **far-field** approximation for the dipole potential:

$$V = \frac{kp}{r^2}\cos\theta \tag{23.12}$$

where p is the dipole moment equal to qd and shows that the potential drops with the square of the distance from the dipole.

The simplest calculation of potential is inside a capacitor because the field is constant. Using equations (23.8) and (23.6), the voltage difference across the capacitor is simply

$$\Delta V = Qd/\epsilon_0 A$$

Though an insulator can introduce leakage into the capacitor, it can also dramatically increase the ability of the capacitor to store charge. This is due to the fact that the charges on the capacitor plates induce dipole moments in the insulator (called the **dielectric**) which effectively reduces the distance between the positive and negative charges. So with the same voltage difference, the capacitor will hold a lot more charge. This relationship is called its **capacitance**:

$$C = Q/V \tag{23.13}$$

For an ideal capacitor, we have

$$C = \kappa\epsilon_0 A/d \tag{23.14}$$

The symbol κ is called the **dielectric constant** and is the factor of increase due to the dielectric between the capacitor plates.

Charging up a capacitor takes work. Each incremental charge that is added takes more work because it is repelled by the ones that are already there. The work required for each must overcome the potential energy given by equation (23.9). But the voltage is constantly increasing. The total work required to bring a capacitor up to voltage is[10]

$$W = \tfrac{1}{2}CV^2 \tag{23.15}$$

Next week, we will cover two topics: electricity and magnetism. In the present lecture we have only discussed what happens when charges are in equilibrium or at rest. In the following lecture we will discuss the flow of charge, or current.

[9] This is similar to the trick we used to analyze the double slit in Lecture 21.

[10] The factor of one-half occurs here in the same way it occurs in the formula for kinetic energy.

This will allow us to investigate some simple electronic circuits with resistors and capacitors.

In addition, we will introduce the magnetic force and discuss its basic properties. We will find a strong parallel between the electric force and the magnetic force with one significant exception: there is no such thing as a magnetic charge. The apparent parallel will continue to erode as we discover that every flow of current creates a magnetic field. Ultimately, natural magnetism is nothing but the result of electric charges in motion.

Lecture 24

Basic Electronics and Magnetism

Read sections 20.1–20.8, 20.11–20.14 and 21.1–21.8

Every electric circuit is designed to control and harness the flow of electric charge. Electric **current** is defined as the amount of charge that flows through a particular surface area each second:

$$I = \frac{\Delta q}{\Delta t} \tag{24.1}$$

The SI unit for current flow is the **ampere** equal to one coulomb per second. Current is driven by a differential in electric potential or voltage. When a charge is exposed to such a differential, it will accelerate according to Newton's second law (5.1). This is what happens in a cathode ray tube, for example. But normally, the flow of current in an electric circuit is controlled through resistance. This means that current is flowing at some constant terminal velocity. And just to add more confusion, this potential difference is typically called the "electromotive force" or **EMF** although it is technically not a force.

When current flows, energy is moving through the circuit. The power consumed by any electric circuit can be obtained by combining equations (23.9) and (24.1). We get:

$$P = IV \tag{24.2}$$

An ideal battery is an electrical component which is able to maintain a constant voltage difference across its terminals regardless of how much current flows. All batteries wear out, but good ones last, so we will assume that our batteries carry a constant voltage differential. A battery provides direct current which is constant over time. We will consider a different form of current (alternating current) in Lecture 25.

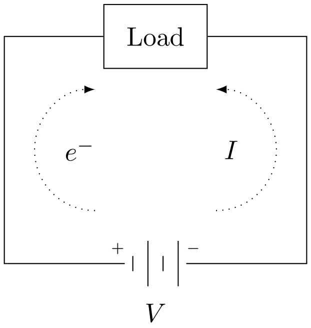


Figure 24.1: A simple electric circuit

Every circuit has a source of voltage difference (a battery, a solar cell, an antenna). The rest of the circuit is called the "load" and is said to "drop" the voltage (see Figure 24.1). Remember that a perfect conductor does not support a voltage difference. So, in order for the circuit to be in a stable state, the voltage differential from the battery must be absorbed in the rest of the circuit somewhere.

Notice that electrons are actually leaving the negative terminal of the battery and traveling to the positive terminal. However, we will ignore this and consider the current to flow from the positive to the negative terminal.[1] Surprisingly, it rarely makes a difference which way you imagine the current to flow.

The simplest electric circuit consists of a battery, a resistor, and two wires connecting them. A **resistor** is specifically designed to drop the voltage and disperse the energy through heat. The ratio of the amount of current that flows per volt

[1] This is just a historical accident because the current carrier has a negative charge. When Franklin established the convention in the 18th century, he guessed wrong.

through a substance is called its **conductivity**, though it is much more common to quote its reciprocal called **resistivity** which may depend on temperature. An ideal resistor has a constant resistance over all voltage ranges and therefore obeys

$$V = IR \tag{24.3}$$

which is called **Ohm's law**. The total resistance of a particular resistor will be a combination of its geometry and the natural resistivity of the material used. We have:

$$R = \rho d/A \tag{24.4}$$

where ρ is the resistivity of the material, d is the length of the resistor, and A is its cross-sectional area.

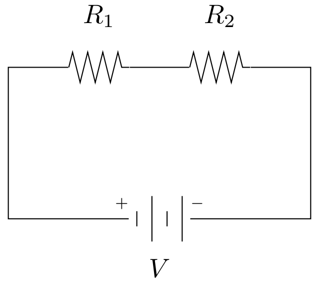


Figure 24.2: A simple series circuit

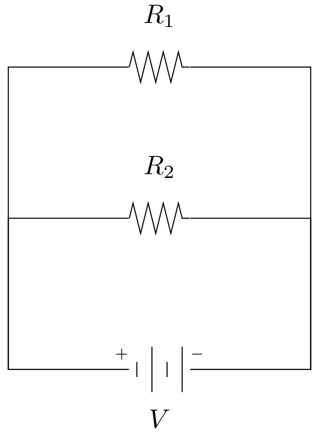


Figure 24.3: A simple parallel circuit

Now suppose we have two resistors connected to a battery as shown in Figure 24.2. These are said to be **in series** because the current must from through one before the other. Clearly, the current through each resistor is equal, and the voltage drop is split between them. The total resistance is simply the sum of the two resistors:

$$R = R_1 + R_2 \tag{24.5}$$

Another way to connect the two resistors is in parallel—see Figure 24.3. In this case, the current runs through the two resistors simultaneously. The current is split and both resistors drop the total voltage. The net resistance of the two parallel resistors is given by:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \tag{24.6}$$

Both equations (24.5) and (24.6) can be seen as consequences of equation (24.4). In the case of resistors in series we effectively increase the length of the resistor: since distance is in the numerator, the resistances add. But in the case of resistors in parallel we effectively increase the area of the resistor and since the area is in the denominator, the reciprocal of the resistances add. Equation (24.6) can also be rearranged as

$$R = \frac{R_1 R_2}{R_1 + R_2}$$

We can add and combine resistors in many combinations. It is even possible to combine them in ways which are neither in series or parallel. The flow of current through any circuit is governed by two basic principles called **Kirchoff's laws**:

- The current at any junction sums to zero. Be careful to take into account the direction of the current.
- The voltage around any loop sums to zero. Be careful to take into account the direction of the potential difference.

The first is based on the conservation of charge and the second on the conservation of energy—compare equation (23.7). For seriously complicated circuits it is sometimes easier to set up the problem using these laws. But typically, we can work the problem by considering the resistors in pairs, either series or parallel.

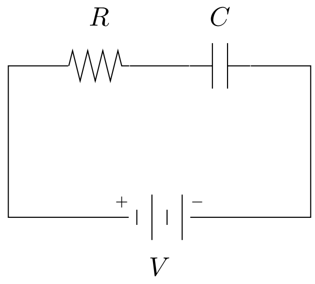


Figure 24.4: A simple RC circuit

We can also add capacitors to our circuits. Consider Figure 24.4. Initially, current flows through the resistor and charges up the capacitor. All the voltage drops across the resistor. But as the capacitor begins to charge up, it also begins to drop voltage according to equation (23.13). Since the voltage drop across the resistor is less, the total current flow is less by Ohm's law (24.3). The flow of current exponentially decreases until the capacitor is essentially fully charged. The formula for the current is given by

$$I(t) = I_0 \exp(-t/RC) \tag{24.7}$$

where $I_0 = V/R$. The quantity RC is sometimes called the **time constant** for the circuit. Typically the capacitor is considered fully charged after five time constants since the current flow is under 1% of its original level.

We are now ready for a complete change of subject. . . . Let's talk magnetism. The awareness of magnets and magnetism appears to go back thousands of years, but the Chinese were the first to write about using magnetic compasses for navigation in the 11th century. The study of magnetism was also one of the few sciences to advance in Europe during the Middle Ages culminating the work of William Gilbert (a contemporary of Galileo) in 1600.[2]

After one has played with a set of bar magnets, it becomes clear that the ends of the magnet are special—we call them the **poles** of the magnet.[3] We say that like poles repel while unlike poles attract. There really is nothing to prevent us from calling one positive and the other negative, but for obvious historical reasons we call these groups north poles and south poles.

Suppose we bring the north poles of two magnets together. They will repel while the south poles attract. This pair of forces will tend to twist the magnet. In other words, each magnet experiences a net torque. This torque guarantees that two magnets will always be attracted to one another, even if they have to twist around first.

Coulomb performed experimentation to quantify this magnetic force. The force law is the same inverse square law that he found for electricity:

$$F = \frac{1}{4\pi\mu_0}\frac{p_1 p_2}{r^2} \tag{24.8}$$

where each p represents the strength of the poles and the constant μ_0 is assigned the value $4\pi \times 10^{-7}$. This constant is called the **magnetic constant**.[4] A pole strength of one **weber** will repel a like pole placed at one meter with a force of $10^7/(4\pi)^2$ newtons. The magnetic field and flux are defined in just the same way as with the electric force.

But if one splits a magnet in two, one gets two magnets, not two poles. The smallest element of magnetism observed physically is not a magnetic pole but a magnetic dipole.[5] In other words, there is no "magnetic charge". This is why there is no such thing as a magnetic conductor nor magnetic circuits. It also implies that Gauss' law for magnetic flux is simply

$$\sum_{\text{surf}} \Phi_{\text{mag}} = 0$$

Though an exact parallel between electricity and magnetism is broken, the two are deeply interconnected. So much so that we are justified in considering them two aspects of a single electromagnetic force. The first step in this discovery is by recognizing that an electric charge can experience a magnetic force. This is called the **Lorentz force**:

$$F = qvB\sin\theta \tag{24.9}$$

where v is the speed of the charge and θ is the angle between the velocity and the magnetic field.[6] Notice that the charge only experiences this force when it is in motion. Also, the direction of this force is perpendicular to both the velocity and the field.

The perpendicular nature of the Lorentz force is unusual, but not unheard of. We encountered a force that acts perpendicular to velocity in the Coriolis force from Lecture 9 (page 52) and in gyroscopic motion. However, knowing that the force is perpendicular to the plane formed by the velocity and field leaves us still uncertain: which way? The rule to distinguish which is called the **right hand rule** and goes like this. Take the four fingers of your right hand and point them

[2] In this work, Gilbert explained the working of the compass needle by assuming the Earth to be a large magnet. He did not know why the Earth is a large magnet, and neither do we.

[3] The poles of a magnet are not actually located at the ends of the magnet. Rather they are somewhat inside (about 1/6 from the end or so).

[4] Again, I learned this as the "permeability of free space", but this is easier to remember

[5] For this reason many authors avoid discussing magnetic poles in detail.

[6] This is exactly the force that is used to guide the electron beam in a standard CRT.

in the direction of the field. Point your thumb in the direction of the charge's motion. If the charge is positive, your palm pushes in the direction of the force.

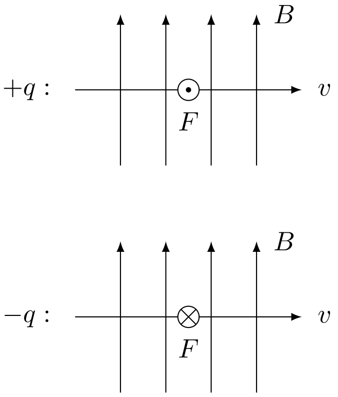


Figure 24.5: The magnetic force on a moving electric charge (Lorentz force)

This weird rule means that magnetic force problems are usually three-dimensional. As such, a convention exists for drawing these things. When we want to indicate the direction of something is pointing out of the paper and toward the reader we use a circular symbol with a dot in the center. If the direction is away from the reader (into the paper), we use a circular symbol with an × in it. These are supposed to represent the tip and tail of an arrow. For example, see Figure 24.5.

Because this force is always perpendicular to the velocity of the charge, the magnetic force never does work on the charge. The force will deflect its motion without changing its speed. In fact, the magnetic force acts as a perfect centripetal force—the typical motion for a charge in a magnetic field is in a circle. Combining equations (24.9) and (6.2), the radius of this circle will be

$$r = \frac{mv}{qB} \tag{24.10}$$

This is how a mass spectrometer works. By sending an ionized sample through a magnetic field, the sample will separate as it bends according to the mass-to-charge ratio (m/q) thus giving important information regarding the constituents of the sample. If we accelerate the sample with a constant electric field, then we can calculate it speed upon entering the magnetic field. Initially, the potential energy of the sample is qV which is converted to kinetic energy $\frac{1}{2}mv^2$. Setting these equal to one another and combining with equation (24.10) yields

$$\frac{m}{q} = \frac{B^2r^2}{2V} \tag{24.11}$$

which shows how the mass-to-charge ratio can be determined from the settings in the mass spectrometer.

Clearly, if a moving charge experiences a magnetic force, so will a straight line of current. Combining equations (24.9) and (24.1) yields

$$F = ILB\sin\theta$$

where L is the length of the wire of current.

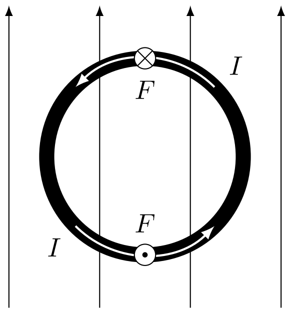


Figure 24.6: How to build a simple motor

If we take the same line of current and wrap it into a circle, the coil will experience a torque in the magnetic field since one side will be pulled while the other side is pushed down. In Figure 24.6, the ring of current is pushed into the paper on the top and out of the paper on the bottom. This force couple puts a total torque on the ring of

$$\tau = NIAB\sin\theta \tag{24.12}$$

where N is the number of coils and A is the area formed by the ring.

This torque is how a basic electric motor works. The magnetic forces tend to twist the coil to align perpendicular to the magnetic field. As it rotates, the torque gets smaller according to $\sin\theta$. Once it is perpendicular, the torque disappears (the magnetic forces are now trying to pull the ring apart). The inertia in the ring will cause it to continue its rotation. If you time it just right and flip the direction of the magnetic field, you can pull the ring back to its original position. By alternating the magnetic field back and forth we can drive the coil around and around indefinitely.

In 1820, Oersted stumbled upon a further connection between electric current and magnetism. Current is not only affected by a magnetic field, but it can also generate a magnetic field. He discovered this by a very simple experiment you can do at home. Take a battery and connect the terminals with a wire (it will get hot, so be careful and quick). Bring the wire close to a typical compass. The compass needle will be deflected in the presence of the wire.

In the end, it was discovered that the magnetic field actually wraps itself around the current. Again, the weird perpendicular nature of magnetism is manifest. And we need another right hand rule to guide us here too. Point the thumb of your right hand in the direction of the current flow. Your fingers naturally curl in the direction of the magnetic field generated by the current.

Shortly afterward, Ampere discovered a general law for the generated magnetic field. **Ampere's law** is

$$\sum_{\text{loop}} B_t \Delta s = \mu_0 I \tag{24.13}$$

which states that the amount that the field curls around the wire is directly proportional to the current in the wire.

Recall in the previous lecture (page 134) that we stated that when a field has zero curl, we may define a field potential. Ampere's law shows that we cannot do the same thing for the magnetic field: there is no magnetic "voltage".

On the other hand, we know that the magnetic field does no work, so the magnetic force is conservative. A type of magnetic potential does exist—but it is a vector (typically given the symbol A). As a consequence, there is no great mathematical advantage to using the magnetic potential, so we will generally ignore it. It does play a role in quantum electrodynamics (see Lecture 27), but we will table the concept for now.[7]

For a simple straight wire, it follows from equation (24.13) that the magnetic field is given by

$$B = \frac{\mu_0 I}{2\pi r} \tag{24.14}$$

This implies that two wires with parallel currents will attract one another. Ampere calculated the total force between the wires to be

$$F = \mu_0 \frac{I_1 I_2}{2\pi r} \tag{24.15}$$

Which looks a bit like Coulomb's law and emphasizes the fact that an electric current acts like the magnetic "charge" we were seeking earlier.

We can drive this analogy a bit further because the magnetic field around a coil of wire is shaped like a dipole field (see Figure 24.7).

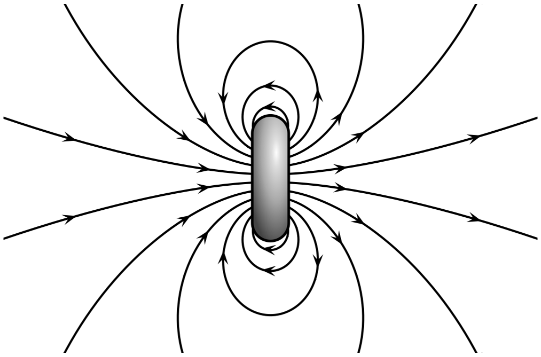


Figure 24.7: Magnetic field around a coil of wire is like a dipole field (image credit here). Compare with Figure 23.4

The dipole moment associated with the current ring is simply

$$m = IA \tag{24.16}$$

where A is the area of the ring. The calculation of force is fairly difficult (like the electric dipole), but the magnetic field in the center of the ring has a magnitude of

$$B = \frac{\mu_0 I}{2r} \tag{24.17}$$

where r is the radius of the coil. Don't confuse this with equation (24.14)!

[7] The vector potential is used to derive the Lorentz force law from the principle of least action. The Lagrangian for a charged particle in an electromagnetic field is

$$\tfrac{1}{2}mv^2 - qV + qv \cdot A$$

The system with the simplest magnetic field is called a **solenoid**. If we take coils of wire and stack them, the magnetic field trapped inside the wires approaches uniformity. This is more commonly known as an **electromagnet**. If we extend the stack infinitely, we have a ideal solenoid and the magnetic field is given by

$$B = \mu_0 n I \tag{24.18}$$

where n is the number of coils per unit length, i.e., $n = N/L$. All of the magnetic field is trapped inside an perfect solenoid and in many ways this offers a magnetic parallel to the electric capacitor.

Ampere's law also works at the subatomic level. Electrons in orbit around the nucleus and even the quantum mechanical spin of elementary particles all have a magnetic moment associated with them. In fact, one of the precision tests of quantum electrodynamics involves the calculation of the magnetic moment of the electron. All this means that each elementary particle and each atom act like little magnets. Typically the orientation of these little magnets is randomized due to thermal agitation. So this natural magnetism is normally hidden at the macroscopic level.

But some materials naturally align their magnetic moments. This is called **ferromagnetism** since iron has this property. The standard bar magnet is made of a ferromagnetic material. Why some materials are ferromagnetic rather than others is a question for solid state physics (see Lecture 28).

There is another way in which large scale objects can be magnetic. As we mentioned earlier, a magnetic dipole in the presence of a magnetic field will tend to align itself with the field. In this way we can induce a dipole moment in an object by bringing it close to another magnet. This is called **paramagnetism** and is similar to the way that static electricity on a balloon will stick to the wall due to an induced electric dipole moment in the wall. In the same way, the effect is typically very slight and difficult to detect.

There is a third way magnetism shows itself on the macroscopic scale called **diamagnetism** but a proper understanding of it requires knowing Lenz' law which will learn about in the next lecture. Unlike the other two, diamagnetism is repulsive so it can create some interesting effects including levitating frogs.

Next week we will continue to explore the way in which electricity and magnetism are intertwined. We will find that we can reverse the way a motor converts electric current into motion. This provides a simple way to generate electricity and is the reason for the prevalence of alternating current in applications. We will explore the way the solenoid works in this context and describe how electrical energy can be transported between separated circuits through the magnetic field. We talk about AC circuits in some detail and finish with a discussion of resonance in electric circuits. This will complete our investigation of electromagnetism with one small but important insight which we will pick up in Lecture 26.

Lecture 25

EM Induction and AC Electronics

Read sections 22.1–22.4, 22.7–22.9 and 23.1–23.4

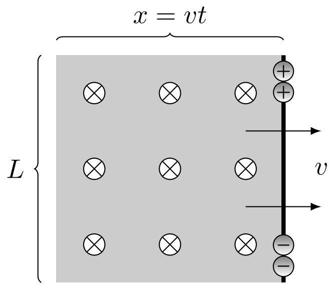


Figure 25.1: Motional EMF generated by an electrodynamic tether

An electrodynamic tether is essentially a huge wire that hangs off a satellite or space craft as it travels through the magnetic field of the earth. The electrons in the conductor are free to move and since they are moving through a magnetic field they are pushed by the Lorentz force. Thus a current is induced in the conductor (see Figure 25.1). This way of driving current is called **motional EMF** and is a simple way (in principle) to extract energy from the earth's magnetic field.

This is an example of a type of electromagnetic induction. We can calculate this motional EMF by blocking the current flow. As the charges separate in the wire, an electric field forms since the charges are attracted to one another (like a capacitor). This electric field counterbalances the Lorentz force caused by the motion of the charges through the magnetic field. At equilibrium the net voltage is zero, so the voltage induced by the charge separation equals the motion EMF causing the charge separation. Setting the electric and magnetic forces equal we have $E = vB$ since the charge cancels from both sides. Using the definition of field potential (23.8), the motional EMF generated in this process must be

$$V_{\text{emf}} = vBL \tag{25.1}$$

where L is the length of the wire.

Equation (25.1) is a special case of **Faraday's law** of induction:[1]

$$V_{\text{emf}} = -\frac{\Delta\Phi_{\text{mag}}}{\Delta t} \tag{25.2}$$

In Figure 25.1, the wire sweeps out the area given by Lvt. The magnetic flux cut by the wire is therefore $LvtB$, which after dividing by t and rearranging gives us equation (25.1).

The negative sign tells us which direction the EMF points and the rule in summarized in **Lenz' law**:

> The magnetic field produced by the current driven by the induced EMF opposes the change in magnetic flux surrounded by the current.

In Figure 25.1, the current flows up the wire which generates a magnetic field out of the paper behind the wire. This field opposes the field in the highlighted area consistent with Lenz' law.

[1] You might wonder why there is no proportionality constant here. The way we have defined the electric and magnetic constants in the previous lectures happens to make the proportionality here equal to one.

Consider another example in which we have a conducting ring sitting perpendicular to a magnetic field (see Figure 25.2). If we shut the field off, current will flow in the wire attempting to restore the field (according to Lenz' law).

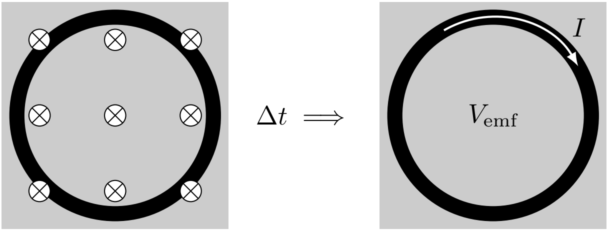


Figure 25.2: Induced EMF in a conducting ring by a changing magnetic field

Now, consider this question: where are the positive and negative terminals of this electromotive force? Answer: there are none—the induced EMF is circular.[2] In other words, when the magnetic field changes, it curls up the electric field!

Perhaps the most important application of Faraday's law is the **electric generator**. In principle, a basic generator is a electric motor operating in reverse—see Figure 24.6. We turn the coil and current flows. Faraday's law operates here because the coil cuts through the magnetic flux. Assuming the coil rotates at a constant angular speed of ω, the flux within a coil with area A is given by

$$\Phi_{\text{mag}} = BA\cos\omega t$$

This assumes the coil is initially perpendicular to the field. Faraday's law (25.2) tells us that the induced EMF in the coil will be[3]

$$V_{\text{emf}} = BA\omega\sin\omega t \tag{25.3}$$

This equation is the reason for **alternating current** in electronics. The sinusoidal pattern of both voltage and current activates Faraday's law and allows us to play with the full power of the electromagnetic field.

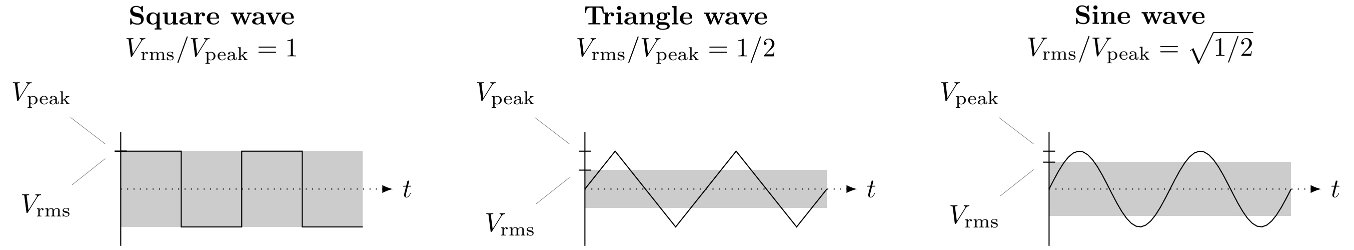


Figure 25.3: Examples of AC voltage patterns and their corresponding RMS values

Technically, alternating current can refer to any repeating pattern of current as in Figure 25.3, but the vast majority of the time we will be considering the sine wave pattern. In the figure I have highlighted the RMS value for the voltage which is a kind of average—a measure of where the voltage is "most of the time". The term RMS stands for **root-mean-square** which is how it is calculated: take the average of the voltage squared then take the square root. This trick eliminates the sign of the voltage—the simple average for all three patterns is zero.

Most of the analysis in our AC electronics will be using the RMS value of the AC voltage. In general, assume the quantities involved are RMS values unless you are otherwise told.

Using alternating current electromagnetic energy can be transmitted between circuits that are not electrically connected. Fluctuating current in one wire will drive current through another wire according to Faraday's law (25.2). This interconnection of two electric circuits through the magnetic field is called their **mutual inductance**. For a wire, the magnetic field drops off according to

[2] This fact perhaps justifies maintaining the distinction between EMF and voltage.

[3] This formula can be verified by using the projection technique we used in Lecture 12 while discussing simple harmonic motion (page 71). It's easiest if you use a square coil. A bit of calculus would be of use here.

equation (24.14) so we have to bring the wires close together before this **electrical interference** manifests itself.

However, we can concentrate this electromagnetic energy by using solenoids instead. This is how the **transformer** works. Suppose we take a solenoid and wrap another on top of it without connecting them. Since the geometries are the same, the magnetic flux in one will be the same as the magnetic flux in the other. The current in the primary coil will drive the magnetic field that influences the voltage in the secondary.

This magnetic field is influenced by the number of coils in the primary according to equation (24.18). The amount of voltage induced in the secondary is related to its number of coils since Faraday's law (25.2) applies to each coil and the EMF accumulates. This means that the voltages are related to the number of coils according to

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \tag{25.4}$$

The ratio on the right-hand-side is called the **turns ratio**. The transformer is said to "step-up" or "step-down" the voltage depending on whether the turns ratio is greater than or less than one, respectively.

Any change in voltage comes at a cost: the currents are inversely proportional to the turns ratio. This fact will conserve the overall energy in the system according to equation (24.2).

Of course, the magnetic field from a current also induces an EMF back onto itself. This EMF actually opposes the original current flow and is called **back EMF**. It's a kind of natural resistance from the electromagnetic field itself and closely related to Lenz' law. What this means is that every circuit has its own **self-inductance**. We define the self-inductance of a circuit through the equation

$$V_{\text{emf}} = L\frac{\Delta I}{\Delta t} \tag{25.5}$$

The self-inductance defines how sensitive the back EMF a circuit is to changes in current. In particular, the self-inductance of an ideal solenoid is

$$L = \mu_0 n^2 A d \tag{25.6}$$

which can be derived by combining equations (24.18), (25.2) and (25.5).

When we talk about a solenoid in an electronic circuit, we use the term **inductor**. Similar to the way a capacitor requires work to come up to full voltage, the inductor requires work to come up to full current since we are working against the back EMF from Faraday's law. The total work required can be calculated from equation (25.5). It is

$$W = \tfrac{1}{2}LI^2 \tag{25.7}$$

This energy is stored in the magnetic field inside the inductor. This parallels the way the energy required to charge a capacitor is stored in its electric field.[4]

So far, we have discussed three basic electronic components: the capacitor, resistor, and inductor. We now wish to investigate how these three components work in AC circuits. But before we do, I would like to introduce you to some mathematical trickery known as **phasors**.

Typically phasors are introduced as complex numbers, but the important thing to know is that they are two-component quantities. Essentially we are going to use the analogy we made in Lecture 12 between uniform circular motion and simple harmonic motion (see Figure 25.4). We represent our oscillating quantity (i.e., voltage and current) with a vector that uniformly rotates around a circle. Its angular speed is $\omega = 2\pi f$.

[4] The parallel is not as exact as it might look. For example, see an interesting discussion on how to release the magnetic energy stored in the inductor (to "discharge" it) here and here.

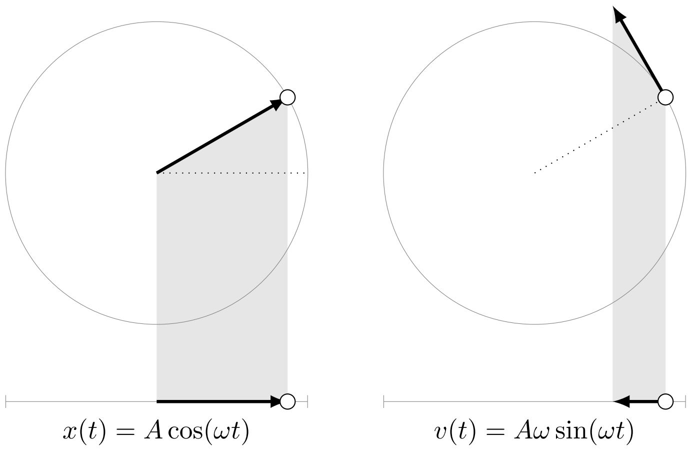


Figure 25.4: Actually, we have already used phasors: taken from Figure 12.2

The most important thing to realize from Figure 25.4 is that if any quantity oscillates according to $A\cos(\omega t)$, then its rate of change also oscillates. The formula for its rate of change is $A\omega\sin(\omega t)$.

For example, if we run an alternating current through an inductor we expect it to generate an EMF due to its self-inductance by (25.5). If the pattern of the current is $I(t) = I_0\cos(2\pi f t)$, then the voltage across the inductor will be

$$V(t) = (2\pi f L)(I_0)\sin(2\pi f t)$$

The ratio of the RMS values of the voltage and current is called the **reactance** of the inductor:

$$X_L = \frac{V_{\rm rms}}{I_{\rm rms}} = 2\pi f L \tag{25.8}$$

This self-inductance means that when an inductor is attached to a high frequency AC voltage source the resulting current will be small since most of the energy will oscillate in the magnetic field.

A capacitor reacts quite differently to high frequency AC voltage. The faster the frequency, the faster the charges have to move to keep up according to equation (23.13). Since current is the flow rate of charge we can rewrite this equation as

$$I = C\frac{\Delta V}{\Delta t}$$

So if the pattern of the voltage is $V(t) = V_0\cos(2\pi f t)$, then the current across the capacitor will be

$$I(t) = (2\pi f C)(V_0)\sin(2\pi f t)$$

Which means the capacitor has a reactance of

$$X_C = \frac{V_{\rm rms}}{I_{\rm rms}} = \frac{1}{2\pi f C} \tag{25.9}$$

Resistors, capacitors, and inductors are called **passive components** because the magnitude of current and voltage are proportional through equations (24.3), (25.9), or (25.8), respectively.[5]

Notice that in the case of resistance the current and voltage wave patterns are in phase. But with reactive components, there is a phase shift. For inductors, the voltage is tied to the change in the current. So the voltage peaks at the initial rise of the current and drops to zero when the current hits its peak. There is a phase shift between the two wave patterns: the voltage is 90° in front of the current.

For a capacitor, the current is tied to the change in voltage. So the current is 90° in front of the current. One way to remember the phase shifts for the two

[5] Active components include diodes and transistors which we will talk about in Lecture 28.

components is with the mnemonic **ELI the ICE man**: voltage (or EMF) leads current in an inductor and current leads voltage in a capacitor.

We distinguish reactance from resistance because the power is lost in resistance. With reactance the power used to maintain either the electric or magnetic field is recovered. Resistance corresponds to friction in the electron flow, while reactance corresponds to its inertia. But they both affect the relationship between voltage and current in the circuit.

Because of the phase shifts involved, calculating the net current flow through a particular AC circuit is a challenge. To facilitate the mathematics we define the **impedance** of a circuit to be the phasor with a magnitude equal to the ratio of voltage and current (like reactance and resistance) and with an angle equal to the phase angle that the voltage leads the current. The impedances for our three electrical components are listed in Table 25.1.

Component	Impedance	
	Mag.	Ang.
Resistor	R	0°
Capacitor	$1/2\pi fC$	$-90°$
Inductor	$2\pi fL$	90°

Table 25.1: Component impedances with their phasor direction

In this way we can summarize equations (24.3), (25.9), and (25.8) in the single AC version:

$$V = IZ \tag{25.10}$$

where Z represents the impedance of the component. This equation is even better than it looks because if we use complex numbers, the phase shift relationships are accounted for automatically.

The phase angle of the total impedance is related to the power consumption in the circuit. The formula is

$$P = IV\cos\phi \tag{25.11}$$

Although this looks a lot like equation (24.2) from which it is derived, the symbols have a slightly different meaning in this AC context. The current and voltage are RMS values and the power is the average power consumed over time. Consider Figure 25.5. It's an odd mathematical coincidence that when two sine waves with a phase shift are multiplied the result is another sine wave which is displaced vertically. This displacement is equal to the average power consumption.

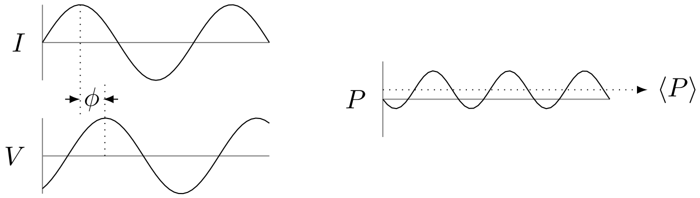
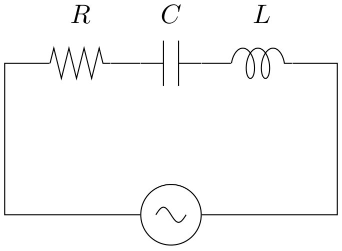


Figure 25.5: How power is consumed in an AC circuit

What is happening is that when the power curve is on the up swing, the energy from the voltage source is being consumed by the resistor and reactance. In other words, some of the energy is going into feeding the electric and magnetic fields. On the down swing, this energy is released back to the circuit though the resistor is still consuming energy. On average we are left with just the impact of the resistor—which is the net consumer of power.

By the way, equation (25.11) shows explicitly that a capacitor and inductor consume no net power because $\cos(90°) = \cos(-90°) = 0$.

Clearly our analysis will be fairly challenging for complex circuits, but we will keep it simple (relatively speaking). Consider the RCL circuit in Figure 25.6. The components are in series so the impedances add. But we must add them like vectors to take into account the phase differences. We will end up with a Figure like 25.7. This diagram assumes a signal frequency of 100 hertz and the RCL values are 33 ohms, 100 microfarads, and 8.2 millihenries respectively.



$$V(t) = V_{\text{rms}}\sqrt{2}\sin(2\pi ft)$$

Figure 25.6: Simple RCL circuit

The total impedance for this series combination is 34.7 Ω with a phase angle of −18°. If the signal voltage (RMS) is 10 volts, the current will be 0.29 amps, and the power consumption will be 2.74 watts.

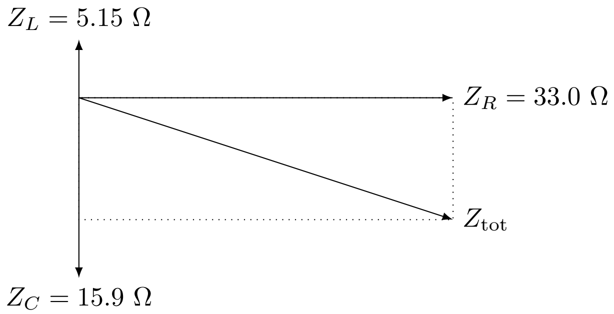
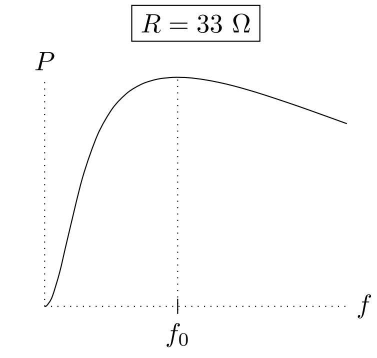


Figure 25.7: Phasor calculation



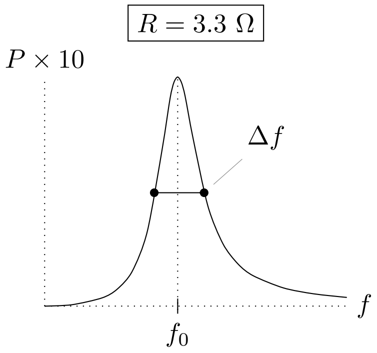


Figure 25.8: Resonance in the RCL circuit

Notice that the voltage across the RCL components are 9.51, 4.59, and 1.48 volts respectively. Clearly these numbers do not add back to the 10 volts from the source. Remember that these individual voltages are the RMS values across the components. If we were to account for the phase differences (the picture would look just like Figure 25.7) we do get back to the original 10 volts.

One important aspect of a circuit like Figure 25.6 is that the current flow is dependent on the frequency of the source voltage. When the frequency is high, the inductor has a high reactance which blocks current. At low frequency, the capacitor has high reactance which blocks the current. In the middle, we have a spike in current flow. This is an example of electrical resonance (see Lecture 12). The inductor acts like the mass, the capacitor acts like the spring, and the resistor is the damping force. Once the current reaches its steady state, the capacitor and inductor require no more power: the energy merely oscillates between the electric and magnetic fields. This leaves the resistor as the only source of impedance. The formula for the resonant frequency of the series RCL circuit is:

$$f_0 = \frac{1}{2\pi}\frac{1}{\sqrt{LC}} \tag{25.12}$$

The phase shift also depends on the frequency. A plot of the power consumption of our circuit from Figure 25.6 is the upper plot of Figure 25.8. If we reduce the resistance by a factor of ten we get the plot underneath it.

The reduction in resistance increases the current flow which increases the overall power consumption of the circuit. But it also sharpens the resonance peak. The range of frequencies for which the power exceed half of its maximum value is called the **bandwidth** of the circuit. For this RCL circuit the bandwidth is

$$\Delta f = R/L \tag{25.13}$$

A better measure of the sharpness of the peak will normalize for the resonant frequency it surrounds. This is called the **quality factor** for the circuit. In this example we have

$$Q = \frac{f_0}{\Delta f} = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{25.14}$$

This Q-factor is important if we are trying to target a particular frequency with our resonant circuit (in a radio, for example).

Next week we will tie up one last loose end in electromagnetic theory. With this last piece, we will see that light is an electromagnetic wave. This will open up the whole non-visible spectrum of frequencies for investigation: radio waves, microwaves, gamma rays, etc. The dark side of this story is that the new theory is in conflict with Newton's laws of motion. Ultimately the resolution is found in Einstein's special theory of relativity. We will end up redefining space, time, mass, momentum and energy before we are done. Finally, we will touch on how relativity theory affects our other long-range force: gravity—which will require the general theory of relativity.

Lecture 26

Electromagnetism and Relativity

Read sections 24.1–24.4 and 28.1–28.7, review Lectures 5, 7, and 8

There is a problem in what we have developed so far—the problem is in Ampere's Law (24.13). It only works if the flow of current is continuous. To see why, consider what happens when we charge a capacitor. If we draw the circle A around the incoming current, Ampere's law tells us that the magnetic field curls around the wire (see Figure 26.1). What Ampere's law does not tell us is which surface to use to determine I. No current pierces the dashed surface indicated. Using this surface, we could conclude that there is no magnetic curl at all.

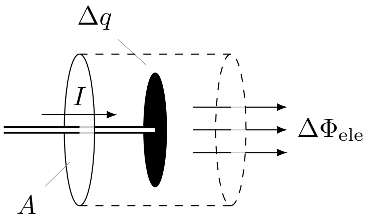


Figure 26.1: Ampere's law needs Maxwell's correction

This inconsistency went unnoticed until Maxwell offered a solution in 1861. Physically, we know that if current is flowing into a region without coming out, charge must be piling up somewhere. In Figure 26.1 the charge is piling up on the plate of the capacitor. Gauss' law (23.4) tells us that the electric flux is increasing too. Although current doesn't pierce the dashed surface, this flux does. So Maxwell added a correction to Ampere's law by assuming that a changing electric flux will also curl the magnetic field:[1]

$$\sum_{\text{loop}} B_t \Delta s = \mu_0 \epsilon_0 \frac{\Delta \Phi_{\text{ele}}}{\Delta t} \tag{26.1}$$

I like to call this **magnetoelectric induction** to emphasize an analogy to Faraday's law (25.2).[2] In Faraday's law we see that a changing magnetic field can create an electric one. In Maxwell's correction (26.1) we see that a changing electric field can create a magnetic one.

Maxwell took this one step further. Prior to Maxwell, physicists had suspected a connection between light and electromagnetism. He was able to show that his additional term implied that the electromagnetic field could act as a medium supporting a wave. The energy of the wave resonates back and forth between the electric field and the magnetic field, like an RCL circuit without the circuitry.

He calculated the speed of this wave as

$$c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} = 2.998 \times 10^8 \tag{26.2}$$

which is the speed of light.[3] So with one minor modification Maxwell completed electromagnetic theory and derived a wave theory of optics. For this, these equations are now collectively called **Maxwell's equations**.

[1] Don't worry: I know this formula is ugly. I only include it for completeness. We won't have to use it in this class.

[2] Maxwell called the term $\epsilon_0 \Delta\Phi_{\text{ele}}/\Delta t$ on the right **displacement current**.

[3] Equation (26.2) shows that Maxwell's correction (26.1) is on the order of $1/c^2$. We will see in a bit that this should be considered a relativistic correction.

In general, the energy density of the electromagnetic field is given by

$$u = \tfrac{1}{2}(\epsilon_0 E^2 + B^2/\mu_0) \tag{26.3}$$

for which equations (23.15) and (25.7) are special cases. Notice how the energy is proportional to the square of the field magnitude similar to (12.9).

But for the special case of electromagnetic waves, the electric and magnetic fields are perpendicular to each other. In addition, the magnitude of the fields are related by $E = cB$. The oscillation of the fields are transverse and they are perpendicular to flow of energy.[4] In fact, the intensity of the radiation is given by[5]

$$I = P/A = cu = \tfrac{1}{2}c\epsilon_0 E^2 \tag{26.4}$$

The factor of one-half is there because the electric field is oscillating. The value of E represents the amplitude or peak value of the wave.

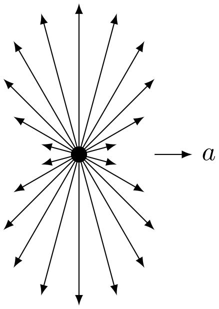


Figure 26.2: Radiation pattern from a charge accelerating to the right (or left)

This electromagnetic radiation occurs whenever a charged particle accelerates. Most of the energy is distributed perpendicular to the direction of the acceleration as in Figure 26.2. The total power radiated is given by **Larmor's formula**:

$$P = \frac{a^2 q^2}{6\pi\epsilon_0 c^3} \tag{26.5}$$

where q is the magnitude of the charge and a is its acceleration. It is this formula which predicts the death spiral of the electron that we mentioned in Lecture 22.

One of the most dramatic consequences of Maxwell's discovery of electromagnetic waves was the opening up of the entire non-visible spectrum. Shortly afterward Hertz confirmed this by generating and capturing radio waves which sit on the lower frequency end of visible light. High energy sources like quasars and nuclear radiation produce EM waves that sit on the higher end of the frequency spectrum.

By equation (26.5) any wire with alternating current will generate EM radiation. Similarly, any wire will react to EM radiation. This is the basic physics behind any wireless technology (radio, TV, router, cell phone, etc.) Typically, the electromagnetic fields associated with EM waves are so feeble that some sort of amplification is required to pick up these signals. This is why sometimes a lightening strike will cause electrical interference. The use of frequency bands are regulated to avoid having people interfere with one another's electronics.

Since the acceleration of the charges is related to the frequency squared (cf. Lectures 25 and 12), the power radiated increases with the fourth power of the AC frequency. This is why the sky is blue. As light from the sun passes through the atmosphere, it causes the atoms of the atmosphere to oscillate back and forth. This motion causes EM waves to re-radiate—perpendicular to the motion of the original radiation. This effect occurs at the highest end of the spectrum most, which is the blue end of the solar spectrum.

Clearly we could go much deeper into the study of electrodynamics, but here we stop. We now turn to discuss special relativity. Although relativity was uncovered through the study of electromagnetism,[6] Einstein saw that the consequences were not just isolated to electromagnetic theory but stuck at the very foundation of Newtonian mechanics.

Most explanations of relativity start with a **light clock** and so will ours. Imagine a thin tube with mirrors on the ends in which a light pulse is bouncing back and forth. Orient it vertically and allow it to move to the right with constant speed. The trajectory of the light pulse as it bounces from bottom to top forms a triangle as in Figure 26.3. This innocuous looking diagram is the Trojan horse of relativity. All of our troubles begin with the following argument.

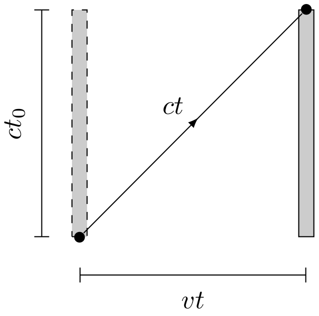


Figure 26.3: Light clock and time dilation

Light travels with constant speed, so the height of the clock is related to the time it takes for light to travel up the clock by ct_0. But if the clock is moving to the

[4] The polarization of the wave is usually defined by the direction of the electric field.

[5] Compare this intensity calculation with that in Lecture 18.

[6] Einstein's paper in 1905 was titled "On the electrodynamics of moving bodies".

right, the trajectory of the light runs up the hypotenuse of the triangle in Figure 26.3. The distance it moves to the right is vt where t is the time it takes the light pulse to hit the top of the clock. Clearly $t > t_0$. In fact, we have

$$(ct)^2 = (ct_0)^2 + (vt)^2 \tag{26.6}$$

which can be rewritten as

$$t = t_0/\sqrt{1 - v^2/c^2} \tag{26.7}$$

Which is called the **time dilation** formula because the moving clock ticks slower.

Hopefully you are saying to yourself, "Slow down—how can such a simple set-up lead to such a sweeping statement about the nature of time itself?" There is one thing that makes this argument special: we are using light. If this were a bullet from a gun, the speed in the vertical direction would be unaffected by the horizontal motion. The two velocities would add (as vectors). This is an application of the principle of relativity: the bullet would take the same amount of time to cross the distance regardless of the horizontal motion.

But this is light—which only moves at the speed of light. Now, the same is true of sound: it travels at a certain speed through to its medium (air). But light is different because it always travels at the speed of light no matter what. This is what makes the Michelson-Morley null result so important. It shows that light is not like a sound wave.

Though many physicists realized that something was up, Einstein saw the heart of the issue. If light speed is the same no matter the frame of reference used, then the principle of relativity forces us to conclude that time is an individual phenomena. My time is not the same as your time—though they are related through equation (26.7).

Suppose we span the distance $L_0 = vt$ with a stick. From the perspective of an observer moving with the clock, this stick is traveling the opposite direction with the same speed so $L = vt_0$. Since the times are related via equation (26.7), the measured lengths are related via

$$L = L_0\sqrt{1 - v^2/c^2} \tag{26.8}$$

Since $L < L_0$, this is called **length contraction**. In relativity, motion shrinks space and slows down time.

An objection may occur to you. If I think the moving ruler is shorter, won't the moving observer think my ruler is longer? This contradicts equation (26.8). The resolution is that the moving observer will not think my measurement is valid due to another relativistic phenomena called **desynchronization**.

Whenever we measure the length of an object in motion, we have to make sure that we mark the edges at the same time. But moving observers disagree on which events are simultaneous. Suppose we flash a light from the exact center of our ruler. We wait for the beams to bounce and return to center. Based on this duration and the speed of light, we calculate the length. But the moving observer sees something like Figure 26.4. The back end of the ruler runs into the light beam at event A. The front end of the ruler is running away and event B occurs later. But I claim that the length of the ruler is the distance between these two points. The moving observer thinks, "Of course you think your ruler is longer—you are cheating. Your ruler is actually shorter by (26.8), but the time lag of vL/c^2 in your measurements overcompensates by a factor $1/(1 - v^2/c^2)$."

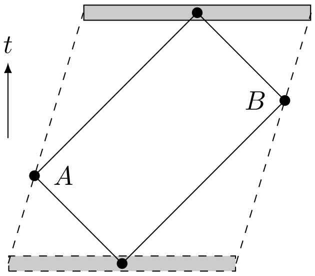


Figure 26.4: Clocks separated in space are desynchronized

All of these effects are summarized in the **Lorentz transformation** which translates the space-time measurements between moving observers. If an observer is moving with velocity v relative to you, her measurements are related to yours according to

$$x' = \gamma(x - vt) \qquad t' = \gamma(t - vx/c^2) \tag{26.9}$$

where $\gamma = 1/\sqrt{1-v^2/c^2}$. This is sometimes called the **Lorentz factor** and is the hallmark of relativity. We can neglect relativistic effects when γ is close to one. Even when v is 10% the speed of light, γ is within 0.5% of one.

These equations are fundamental and affect every other calculation in mechanics you have learned so far. In particular, the velocity of an object as measured in the moving frame will be

$$u' = \frac{\Delta x'}{\Delta t'} = \frac{\gamma(\Delta x - v\Delta t)}{\gamma(\Delta t - v\Delta x/c^2)} = \frac{u-v}{1-uv/c^2}$$

Without relativity, we would simply subtract the motion of the observer from our measurements like $u' = u - v$, but the desynchronization effect in the denominator changes how the velocity measurements interact.

This is also one of the ways we can see why no object can move faster than light. If we start with an object moving at velocity v and increment the velocity by an amount u (less than the speed of light), the final speed is given by

$$u' = \frac{u+v}{1+uv/c^2} \tag{26.10}$$

which will always be less than light speed.[7]

One hidden consequence of equations (26.9) is that between any two events, the combined quantity

$$I = (c\Delta t)^2 - \Delta x^2 \tag{26.11}$$

is the same for both observers.[8] This **space-time interval** is truly the best representation of how space and time are united and intertwined in relativity theory. Investigating the geometry associated with this interval offers a way of viewing the theory that is illuminating and mitigates the feeling of a cosmic conspiracy that time dilation, length contraction and desynchronization generate.

Since velocities don't work the way we are used to in relativity, is it any surprise that momentum and energy don't either? The most egregious situation is that the conservation of our friend mv is no longer invariant. For example, if the total momentum is zero in the center of mass frame, it will not be conserved in the lab frame because of the way velocities combine in equation (26.10).

Fortunately, it can be shown that the conservation of γmv is invariant.[9] This is called the **relativistic momentum** and clearly reduces to the classical definition of momentum when $v \ll c$. However, in order for relativistic momentum to be conserved we must also have γm conserved. This may seem odd until one runs through the classical version of this argument and notices that it requires m to be conserved. So in order for this relativistic argument to work we must redefine both momentum and mass.

The quantity γm is frequently called **relativistic mass**, though this usage has gone out of favor.[10] The main reason is that guessing at a relativistic formula by simply replacing m with γm in most classical formulas won't work. The preferred approach is to multiply this quantity by c^2 and call it **relativistic energy** and use m only to refer to the mass of an object measured from rest (i.e., its **rest mass**). Since γmc^2 is in units of energy, we are okay to call it this but there is an even better reason.

Using the binomial theorem (1.3) we have

$$\gamma = (1-v^2/c^2)^{-1/2} \approx 1 + \tfrac{1}{2}v^2/c^2 \tag{26.12}$$

This is a useful approximation to be aware of it its own right. But is also means we can say

$$E = \gamma mc^2 = mc^2 + \tfrac{1}{2}mv^2 + \ldots$$

[7] Here's the proof. Both u and v are less than c. Thus $c-u$ and $c-v$ are both greater than zero and so is their product. Rearranging this inequality we can say

$$c^2(1+uv/c^2) > c(u+v)$$

Rearrange again and the conclusion follows.

[8] The converse is also true: equations (26.9) are the only linear combinations which preserve this quantity.

[9] The proof is straightforward, but a bit laborious. You will need to know that

$$\gamma_{v'} = \gamma_u \gamma_v [1 + uv/c^2]$$

[10] Although in his physics lectures, Feynman makes the claim that all of relativity can be derived from this one formula. Actually, the idea of relativistic mass goes back to Einstein himself. It's pretty easy to start a flame war on physics related websites nowadays just by bringing up the topic.

The first term is a constant and is called the **rest energy** of the particle since it doesn't go away even if the particle is at rest. The second term is our friend kinetic energy. The remaining terms are relativistic corrections to this kinetic energy formula.[11] A second verification of this approach is possible by calculating the work required to relativistically accelerate a particle from rest. If we use $p = \gamma mv$, then the total work done on the particle is[12]

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

which is the same way we derived the $\frac{1}{2}mv^2$ formula in classical mechanics.

We continue to call a collision elastic if it conserves kinetic energy. But notice that in relativity the total energy γmc^2 is conserved even in inelastic collisions.

Consider two identical particles that are flying toward one another with equal and opposite velocities. Suppose they each are moving fast enough to have a Lorentz factor of $\frac{3}{2}$ so the total energy of the system is $3mc^2$. If they collide and stick (completely inelastic collision), the composite still has this energy. But now it is at rest (net momentum is zero) with a Lorentz factor is one. It follows that the rest mass of the composite is $3m$—the mass of the object continues to hold the incoming kinetic energy.

The reverse of this process works too. This connection between mass and energy is completely unexpected in classical mechanics and is one of the reasons why $E = mc^2$ is so famous. We can calculate the energy locked in the nucleus of the atom by comparing the mass of the composite atom with the mass of its parts. Multiply this difference by c^2 and the number of atoms in a kilogram of uranium and you end up with a very big number.

Occasionally in relativity problems you know the energy and momentum of a particle and need to determine the velocity. One shortcut to use is

$$Ev = pc^2 \tag{26.13}$$

which completely skips the need to deal with γ in the calculation.[13] Another useful connection between these concepts is

$$E^2 - p^2c^2 = m^2c^4 \tag{26.14}$$

This equation is more important than it looks. This is because it parallels equation (26.11) with energy connected to time and momentum connected to space. If we were to dive down the rabbit-hole, we would combine energy and momentum into a single four-vector called **four-momentum**. Similar to the way distance and duration are "projections" of the space-time interval, the energy and momentum of a particle are the "projections" of this relativistic four-momentum. They are dilated, contracted, and mixed via the Lorentz transformation (26.9) just like space and time.

Though we have covered a lot, one obvious gap is left: Newton's second law. The gap is not as critical as it seems because most applications of relativity involve the analysis of collisions for which energy and momentum are sufficient. On the other hand, the means of relativistically correcting a particular force law is not an easy problem.

But we already know that electromagnetism is a relativistic theory. One insight that relativity brings to electromagnetism is that the electric force and the magnetic force are not just intertwined, but can be considered as "projections" of a four-dimensional electromagnetic force. In a similar way to how energy and momentum are "shadows" of a relativistic four-momentum, the electric and magnetic forces are the "shadows" of electromagnetism on time and space, respectively.[14] Though a bit more complicated because we are talking about combining two vectors into some four-dimensional tensor, this approach radically simplifies the equations of electromagnetism.

[11] If you are interested, the next term happens to be $\frac{3}{8}mv^4/c^2$.

[12] Unfortunately, I know of no non-calculus technique of deriving this equation. If you think of one, let me know!

[13] This implies that the energy and momentum of a light pulse are related through $E = pc$. Plugging this into the next equation (26.14) yields $m = 0$. This is why we say a photon has zero rest mass even though a photon is never actually at rest. Reversing the argument works too, so anything with "zero rest mass" can only move with the speed of light.

[14] The two forces also mix together in formulas similar to the Lorentz transformation (26.9). This implies that a force we consider magnetic in one frame may be considered to be electric in another. This fact was what Einstein originally considered as the main theoretical support for relativity.

We have already made the point that fundamentally all the other macroscopic forces (except gravity) are electromagnetic in origin. So in principle, all that's left is a lot of calculations.[15]

Then there is gravity. One key element of any relativistic force law is that its speed of influence cannot exceed the speed of light. So we know right away that Newton's law of gravity (6.3) is wrong. Einstein had the fundamental physical insight in 1907 to solve the problem. It took him another eight years to struggle through the math and find the correct theory which we call **general relativity**.[16] This was an experience he never forgot and is why he wrote the following statement in a 1943 letter that may seem incredible to you:

> Do not worry about your difficulties in mathematics. I can assure you mine are still greater.

Einstein's physical insight is now called the **equivalence principle** and is the simple recognition that a reference frame accelerating under the influence of gravity (in free-fall) is inertial. Mathematically this is because the mass in Newton's second law (5.1) is the same as the mass in the law of gravity (6.3). So everything accelerates together in tandem. Einstein saw that this special property allows one to generalize the principle of relativity to include frames in free-fall.

Here's the trick: we know how to correct Newton's laws in the freely falling frame—use special relativity. What we need to know is how to transform these corrections out of the free-fall frame to one on the ground. The answer is to replace the space-time interval in equation (26.11) with the **Schwarzschild metric**:[17]

$$I(r) = \left(\frac{r - r_s}{r}\right)(c\Delta t)^2 - \left(\frac{r}{r - r_s}\right)\Delta x^2 \qquad (26.15)$$

where $r_s = 2GM/c^2$ and is called the **Schwarzschild radius**. This length is the characteristic distance in which general relativistic effects are important. For the Sun it is three kilometers which is negligible on an astronomic scale. The Schwarzschild radius represents the distance at which light is captured by an object and therefore represents the effective radius of a **black hole**. A black hole need not be massive: it must be very dense. If we could compress the mass of the Earth below a radius of nine millimeters, general relativity would take over and it would collapse into a black hole.

For any stationary clock $\Delta x = 0$ and the space-time interval equals $(ct_0)^2$ where t_0 is the proper time measured by the clock. Because of the factors in the Schwarzschild metric (26.15), a stationary clock in a gravitational well will tick slower than one in deep space (**gravitational time dilation**). We have

$$t_0 = t\sqrt{1 - \frac{2GM}{rc^2}} \approx t - \frac{GM}{rc^2} \qquad (26.16)$$

This effect is directly confirmed in the delay of radar beams bounced off Venus and in gravitational red-shift. It can also be used to see that the speed of light appears to slows down as it approaches the Schwarzschild radius of a black hole.[18] This change in speed can deflect light (like refraction through variable media—see Lecture 20). This prediction and its dramatic confirmation in the eclipse of 1919 launched Einstein to world-wide fame.

I've left a lot out (obviously) including gravitational waves, cosmology, and how this is all related to "curved spacetime". For a few more insights review the end of Lecture 11 and 13, but here we must stop.

Next week we make a complete shift in gears to talk about quantum mechanics. We have already discussed (Lectures 16 and 22) how cracks in our understanding of the very small began to appear around 1900. We will talk about how these

[15] This is extremely unfair, of course. The complexities of how electromagnetism manifests itself as elastic, friction, tension, and other forces depends upon a perfect understanding of the nature of matter. Until that project is done, this statement is untrue.

[16] This branch of mathematics is now known as differential geometry.

[17] Confession: This is the basic idea, but I've cheated here. We need to take into account the polar coordinates used to derive this formula by replacing x with r and we need to add a couple of terms. For the full meal deal see Wikipedia.

[18] Any object falling into a black hole will appear to slow down. In fact it will never appear to cross the Schwarzschild radius. From the object's point of view there is nothing unusual—it just falls, but the outside world never sees anything after this line is crossed. This is why this radius is also called the **event horizon** of the black hole.

issues were resolved over the following 50 years through an understanding of the wave nature of matter. We will also touch on the difficulties and the ultimate success of creating a quantum theory of electrodynamics.

Lecture 27

Quantum Mechanics

Read sections 29.1–29.6 and review Lecture 22

We've already discussed some of the reasons why physicists around 1900 were beginning to witness the breakdown of classical physics. Shortly after relativity made us adjust our understanding of the fast and energetic, a much larger attack was made on our understanding of the very small. If you thought relativity was mind-blowing, you might want to sit down now.

In Lecture 22 we discussed how Planck and Einstein were driven by experiment to postulate a particle theory of light (in flat contradiction to classical electromagnetism). The energy relationship for the photon is given in equation (22.2) as $E = hf$ where f is the frequency of the light. Einstein showed how this relationship could make sense out of the ultraviolet catastrophe and the photoelectric effect.

The next step down the road to quantum mechanics was taken in 1924 by Louis de Broglie. His argument was basically: if photons have both wave and particle properties, then maybe electrons do too. Since the momentum of a photon is given by $p = E/c$ and $f\lambda = c$ for any wave, we have $p = h/\lambda$ for the photon. He postulated that the electron (and any other material particle) must have an associated wavelength of

$$\lambda = h/p \tag{27.1}$$

This is known as the **de Broglie wavelength** of the particle and characterizes its **matter wave**.[1]

At the time, understanding the structure of the atom was the hot topic of the day. De Broglie's hypothesis made sense out of some of the results seen at the time. We will return to this topic in the next lecture. More direct evidence for the wave nature of matter is available now in neutron diffraction experiments and the electron double-slit experiment.

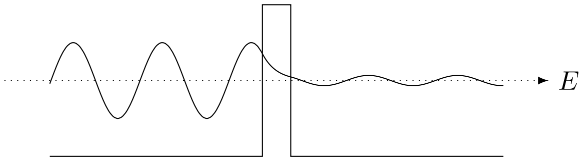


Figure 27.1: Quantum tunneling through a potential energy barrier

One important consequence of the wave nature of matter is **quantum tunneling**. Consider Figure 27.1. A classical particle with kinetic energy E approaches a potential energy barrier of twice this energy. Classically, we expect the particle to "bounce" off of this barrier. However, a wave will propagate into such a barrier though with exponentially decreasing amplitude we called this "attenuation" in

[1] Though "matter field" might be a more appropriate term.

Lecture 21. If the barrier is small enough, there may be enough amplitude left for a small amount of this wave to be transmitted through the barrier.

This is the explanation behind natural radioactivity. The protons are bound together by the strong nuclear force (see Lecture 29) which creates a potential well for the nucleus. But outside the nuclear range, the protons are repelled by their electrostatic repulsion. This combination of the forces produces a barrier through which, occasionally, some of the nuclear content escapes.

Overall, De Broglie's theory is suggestive but incomplete. The **Schrödinger wave equation** published in 1926 is what we seek:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = (E-V)\psi \tag{27.2}$$

where $\hbar = h/2\pi$. This form of the equation assumes the potential energy V is constant so the physical system is in a standing wave mode—like a stable atom. The one-dimensional solution to this equation is

$$\psi = A\cos(\pm kx+\phi) \quad \text{with} \quad \hbar k = \sqrt{2m(E-V)} \tag{27.3}$$

unless $E < V$ in which case the solution is

$$\psi = A\exp(\pm kx+\phi) \quad \text{with} \quad \hbar k = \sqrt{2m(V-E)} \tag{27.4}$$

These are the formulas used to generate Figure 27.1. These standing waves oscillate with a frequency $f = E/h$.

By its very nature, a wave is extended in space. It is possible to "localize" these waves by combining them with different but similar wavelengths (similar to the Fourier analysis in Lecture 19). This means that for waves there is an inverse relationship between the position and wavelength. But the wavelength of our matter waves is connected to the momentum of the particle via (27.1). The **Heisenberg uncertainty principle** summarizes this relationship as

$$(\Delta x)(\Delta p) \geq \hbar/2 \tag{27.5}$$

The uncertainty principle shows that identifying the exact trajectory of a quantum particle is doomed to failure.[2] The more we localize the particle, the greater the uncertainty in its momentum. In other words, the more we know about where it is, the less we know about where it will be.

So, in quantum mechanics the matter wave is primary. But what exactly is ψ? Some insight into this question can be gained by revisiting the photoelectric effect. The intensity of the radiation is

$$I = nhf/t$$

This is simply the total energy per photon (hf) multiplied by the number of photons per second (n/t). This is the particle viewpoint. From the standpoint of electromagnetism, the intensity is given by equation (26.4). Setting these two equations equal to each another yields

$$nhf/t = \tfrac{1}{2}c\epsilon_0 E^2$$

Remember that E in this case represents the peak value of the electric field. As the photon is related to the electromagnetic field, so the electron is related to its matter wave. In other words, it seems reasonable to assume that the square of the wave amplitude is proportional to number of particles present. If we are talking about a single particle, then this value must represent the probability of finding the particle at that location in an experiment. In symbols,

$$\rho \propto \psi^2 \tag{27.6}$$

[2] Although there is an interesting approach to quantum mechanics by David Bohm which claims to do just that.

where ρ is the probability density of locating the particle at this point in space. Of course, the sum of all these probabilities must be one.

We will explore how the matter wave works in the context of the atom in the next lecture. We will also get a hint of how quantum mechanics provides physical foundation for chemistry and solid state physics. But before we do, an obvious question has probably occurred to you. What about relativity? Oddly enough, there are significant conceptual difficulties in building a relativistic quantum theory. It took physicists 20 years to untangle this Gordian knot and a trio of men (Tomonaga, Schwinger, and Feynman) received the Nobel Prize in 1965 for it.[3]

Quantum field theory (QFT) is the result of combining quantum mechanics and special relativity. In both there is a standard recipe for "graduating" or generalizing our classical models.

In special relativity we replace our notions of space and time with space-time and recognize that both distance and duration are merely components of the space-time interval between the two events. In addition, pairs of mechanical quantities are combined in an analogous way—in particular, energy and momentum are related according to $E^2 = p^2c^2 + m^2c^4$ (26.14) which replaces the Newtonian relationship $KE = p^2/2m$.

In quantum mechanics we replace the notion of an elementary particle with its "wave-function" $\psi(x)$ which can be used to calculate all of the mechanical quantities related to this quantum particle. In particular, the momentum of the particle is related to the slope of the wave and the energy is related to its rate of change. In symbols:

$$p(\psi) = -i\hbar\frac{\partial\psi}{\partial x} \qquad E(\psi) = i\hbar\frac{\partial\psi}{\partial t} \tag{27.7}$$

Starting from

$$p^2/2m = KE = E - V$$

yields Schrödinger's equation (27.2).

The most natural combination of these two recipes starts with the relativistic equation (26.14) and imposes the quantum recipe to yield

$$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \frac{\partial^2\psi}{\partial x^2} + \frac{m^2c^2}{\hbar^2}\psi = 0 \tag{27.8}$$

which is now known as the **Klein-Gordon** wave equation. However, this equation can be shown to yield negative probabilities for the location of the quantum particle. This makes interpreting this probability density problematic.

The correct conclusion is that there is no such thing as a quantum field theory for a single particle.[4] The reality is that the analysis of any high energy system must take into account the possibility of secondary creation and annihilation events. In quantum field theory, every problem is a many-body problem. We now know how to deal with that (given some reservations)—we have Feynman diagrams.

A typical Feynman diagram looks like Figure 27.2. This is a space-time diagram with time flowing up the page. The lines represent electrons and the squiggly line is a photon. The black dots at the beginning and end are the actual events measured in the lab. The white dots in the middle are called "virtual" in the sense that they are not actually observed—the photon exchange is unseen so it is called a **virtual photon**. In quantum field theory, the electromagnetic force is mediated by the exchange of these virtual photons.

Quantum electrodynamics (QED) is the application of quantum field theory to the electromagnetic force. The photon is governed by Maxwell's equations and the electron is described by the formula discovered by Dirac in 1928. Feynman

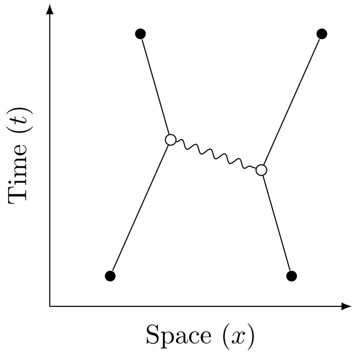


Figure 27.2: A typical Feynman diagram

[3] Feynman's Nobel Lecture is an interesting recounting of some of his journey to unlocking these mysteries.

[4] It seems almost irreverent to skip Dirac's theory of the electron here. But it is truly a logical sidebar to our main point. In a way, Dirac exploits a loophole in the previous logic which delays but in the end does not deny the final conclusion. Of course we need Dirac's electron (i.e., the electron is a spinor field), but not as a motivation for QFT.

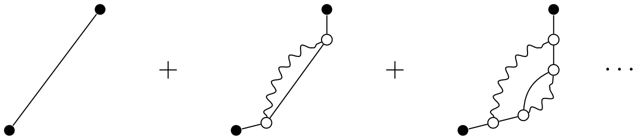
calls it the "crown jewel" of physics and it is the most accurate physical theory in history—confirmed by experiment within ten parts per billion.

The nice thing about these Feynman diagrams is that you can actually see the electrons exchanging the photon. In fact, the diagram is so suggestive, it is important to emphasize what it is not. It does not represent the actual interaction between the electrons—it is a book-keeping tool used to calculate the interaction. The diagram represents a whole class of similar diagrams with the unobserved virtual events located in different places. Truly anything goes as long as the initial and final events are the same.

Each Feynman diagram is used to calculate a complex number with an angle proportional to the total action of the diagram (using the Lagrangian from classical mechanics). Those diagrams representing situations far from the classical prediction typically cancel out. In this way, Feynman found justification for the classical principle of least action. The diagrams also show us a practical method of correcting our classical predictions by including those diagrams close to the classical result.

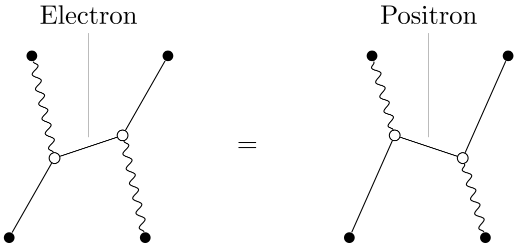
This correction is more difficult than it may appear. Simply calculating the probability involved in one electron traveling from point A to point B involves an infinite number of diagrams (see Figure 27.3).

Figure 27.3: The foremost rule of quantum field theory: anything goes as long as you don't get caught



This shows that the bare electron is always surrounded by a cloud of virtual particles like groupies. The electron we measure in the lab is actually this cloud—one never sees a naked electron.

Figure 27.4: Anything goes—including traveling backward in time



One surprisingly simple consequence of these diagrams is that it is easy to see how QFT predicts the existence of anti-matter. Consider the pair of diagrams in Figure 27.4. The one on the left represents an electron that emits a photon then later absorbs and incoming one. But this is equivalent to the other diagram on the right with this order reversed. But what is going on in the middle? The electron is going backward in time.

From a technical standpoint, these two diagrams are indistinguishable. You can't have one without the other. Feynman interpreted the time-traveling electron as the positron.[5] Look at the second diagram a bit closer. The incoming photon on the right splits into two particles—the one on the right is the outgoing electron. The one on the left is the backward traveling electron. But in the lab we would call this the creation of a matter-antimatter pair, one electron and one positron. Later in time, the positron strikes and annihilates the incoming electron. This leaves the outgoing photon as the byproduct. In this way Feynman explained the existence of antimatter and why every particle has an antimatter twin with the same mass. Simple.

In fact, we can even draw Figure 27.5. This diagram represents the virtual creation then annihilation of an electron-positron pair. Note that there are no incoming

[5] The positron was discovered by Carl Anderson in 1932. It has all the properties of the electron except with a positive charge.

or outgoing particles. This diagram shows that even the vacuum is full of a sea of virtual matter-antimatter pairs. This is the source of the so-called **zero-point energy** of the vacuum. In quantum field theory even the vacuum is a many-body problem!

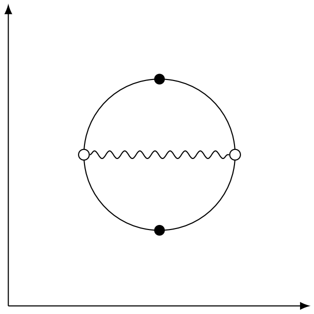


Figure 27.5: An oyster diagram—the vacuum is not empty

One of the most important consequences of quantum field theory is that every quantum particle falls into one of two categories: **fermions** or **bosons**. Bosons are the exchange particles in our Feynman diagrams: they can be absorbed and emitted by fermions and mediate the interaction between them. Fermions represent the "hard" particles which form into atoms and all the matter we see. Fermions obey the **Pauli exclusion principle** which states that no two identical fermions can exist in the same quantum state. There is a natural "repulsion" between them which gives matter its stability. On the other hand, bosons are gregarious: they have a propensity to collect into the same state. This collective behavior is why they appear classically as force fields since they work together to produce a net interaction measurable on a macroscopic scale.

This also helps to explain how a **laser** works. In a typical laser the medium is designed to support the emission of photons. Usually this is done by "pumping" electrons into a high energy state. The electrons each radiate this energy as a photon in order to return to a lower energy state. The lasing material must be able to support the electrons in this unstable position for a relatively long time frame (a few nanoseconds).

The laser begins with one photon. The presence of this photon induces the next photon to be aligned with itself because of this "gregarious" nature of bosons. Put a couple of mirrors on the end and as long as there are electrons to generate the photons, the pattern continues and multiplies. Eventually you have a significant beam of light in which all the photon waves are aligned in both phase and polarization. Since these waves constructively interfere, the total energy content of the beam scales with square of the number of photons. With classical particles the energy scales with only the number of photons. This coherence is why lasers are so much more powerful than other sources of light.

Occasionally it is possible to create a situation in which fermions lock together and act like bosons. This is what happens in **superconductivity**, for example. The electrons pair together by interacting through the metal substrate. This allows the wave function of the electron pair to constructively interfere like bosons. As a consequence, the flow of the electron pairs is self-sustaining—strong enough to overcome the natural resistance of the wire at very low temperatures.

The strange behavior of liquid helium is another example. Helium liquefies at 4.2 kelvin, but at 2.17 kelvin the viscosity of the fluid drops to zero which allows the fluid to flow without resistance. So much so that the capillary action of surface tension will cause the fluid to "wick" up the sides of its container. Every liquid does this, but without viscosity liquid helium crawls up and out of the container completely. This is an example of a **superfluid**.

Usually these quantum phenomena require a very low temperature to manifest because the random motion associated with heat destroys the coherence required to create these weird effects. But in 1986, the first "high-temperature" material ($\mathrm{YBa_2Cu_3O_7}$) was discovered with a superconducting temperature above the boiling point of liquid nitrogen (77 kelvin). This made the practical use of superconductors in electronics possible.

Next lecture we will investigate how these quantum principles explain the structure of the atom. We will see that by trapping the electron field in the electrostatic potential of the nucleus it collapses into stable standing wave patterns. These electron orbitals explain the spectroscopic patterns of the elements and the Pauli exclusion principle explains the structure of the periodic table. We will then cover how solid state electronic devices work.

Lecture 28

Atoms and Solid State Physics

Read sections 30.1–30.6, review section 23.5, see also Lecture 22

As was mentioned in Lecture 27, the primary motivation behind the early development of quantum mechanics was to understand how the atom works. Atomic theory was developed by the chemists of the early 19th century: primarily Dalton and Avagadro. In 1869 Mendeleev compiled the elements into the periodic table in a way which is ordered by atomic mass but also emphasizes their chemical similarity. The periodic table gave order to chemistry, but the reason behind the order was unknown.

In this same time frame, it was discovered that each element has its own characteristic emission spectrum. When burned or electrically excited, the atoms emit electromagnetic radiation at very specific frequencies. These spectral lines act as a kind of fingerprint for each element and this spectral analysis offers a unique way to identify the presence of elements in an otherwise unknown sample. The process works in reverse too: an otherwise continuous spectrum will possess dark lines from the absorption of light at these same frequencies. This is how the composition of distant stars and planets are analyzed.

In the late 1880's a formula for the spectral lines of hydrogen was empirically discovered:

$$\frac{1}{\lambda} = R\left(\frac{1}{n_1} - \frac{1}{n_2}\right) \tag{28.1}$$

where R, the Rydberg constant, is equal to 1.097×10^7 and is the most accurately measured physical constant. There are four lines that sit in the visible range for hydrogen with $n_1 = 2$ and $n_2 = 3$, 4, 5, and 6. These are called the **Balmer lines** and are useful to detect the presence of hydrogen gas in interstellar space.

Then in 1897, the electron was discovered—the first subatomic particle—showing that the atom was not in fact atomic ("atom" in Greek means indivisible). Four years later, Rutherford showed that the bulk of the atom is confined to a central nucleus. Presumably the electron orbits this nucleus like a planet around the sun.

But this planetary model can't work. Using Larmor's formula (26.5) it is possible to show that the electron will radiate all of its kinetic energy in one-hundredth of a nanosecond.

This contradiction remained unresolved until Bohr proposed a novel solution in 1913. Inspired by Planck's quantum formula for radiation, he showed that if we postulate that the orbits of the electron are quantized, we can have a stable atom and explain the spectral lines of hydrogen.

Bohr's assumed that the angular momentum of the electron orbit obeys the relation

$$L = mvr = n\hbar \tag{28.2}$$

Using this relationship and by setting the centripetal force of the orbit equal to Coulomb's law (23.1), one can determine that the radius of the allowed orbits are

$$r_n = \frac{n^2\hbar^2}{ke^2m} \tag{28.3}$$

where e and m are the charge and mass of the electron, respectively. The minimum radius for $n = 1$ is 0.529×10^{-10} meters, which is called the **Bohr radius**. It also follows that the energy of these orbits is given by

$$E_n = -\frac{ke^2}{2r_n} = -\frac{13.6 \text{ eV}}{n^2} \tag{28.4}$$

Furthermore Bohr assumed that each transition between these orbits is accompanied by the emission or absorption of a photon with energy equal to $\Delta E = hf$. From this, the Rydberg formula (28.1) follows.

So with one simple condition Bohr was able to explain a large swath of atomic theory. But why this quantization? This is where de Broglie's matter wave comes in. Each of Bohr's stable orbits correspond to a standing wave similar to harmonics on a string (see Lecture 19).

But this picture is too simplistic. We ought really to consider these standing waves in three dimensions (as in Figure 28.1). The solutions to Schrödinger's equation (27.2) with stable energy are the atomic orbitals. These patterns are characterized by three integers n, ℓ, and m which define their shape. The first integer n is called the **principal quantum number** and is related to the energy of the electron. The larger the n value, the more energy, and the larger the size of the orbital—in the same way as in the Bohr model.

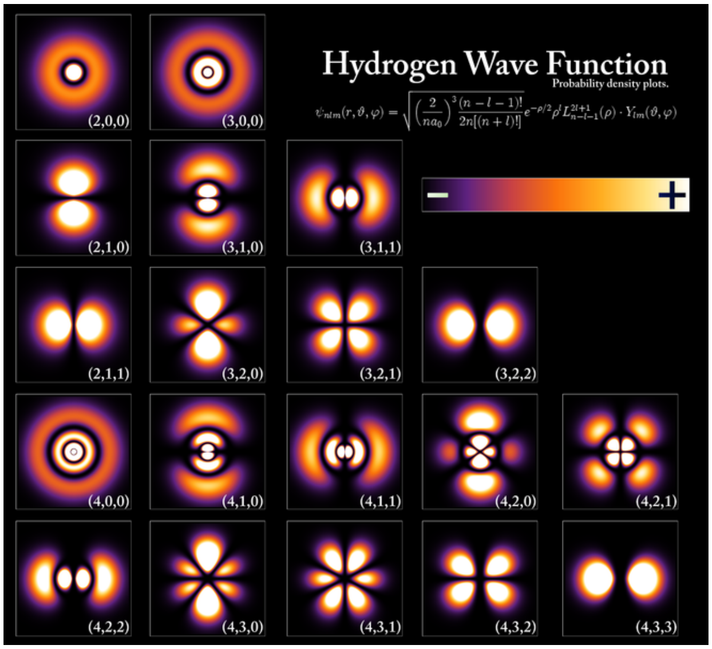

Hydrogen Wave Function
Probability density plots.
$\psi_{nlm}(r,\vartheta,\varphi) = \sqrt{\left(\frac{2}{na_0}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]}} e^{-\rho/2}\rho^l L^{2l+1}_{n-l-1}(\rho) \cdot Y_{lm}(\vartheta,\varphi)$
−
+
(2,0,0)
(3,0,0)
(2,1,0)
(3,1,0)
(3,1,1)
(2,1,1)
(3,2,0)
(3,2,1)
(3,2,2)
(4,0,0)
(4,1,0)
(4,1,1)
(4,2,0)
(4,2,1)
(4,2,2)
(4,3,0)
(4,3,1)
(4,3,2)
(4,3,3)

Figure 28.1: Atomic orbitals of the hydrogen atom (image credit here)

The other two quantum numbers are related to the shape and spin of the orbital. Since they have the same energy we call them **degenerate states**. Because of the way the math works, there are more degenerate states the larger the value of n. The possible values are summarized in Table 28.1.

Principal	n	$1, 2, 3, \ldots$
Angular	ℓ	$0, 1, 2, \ldots, (n-1)$
Magnetic	m	$0, \pm1, \pm2, \ldots, \pm\ell$

Table 28.1: Quantum numbers for an electron trapped in a Coulomb spherical well

For example, the $n = 2$ energy state has a four-fold degeneracy because ℓ can be either zero or one. If zero, m must have the value zero. If $\ell = 1$, then m can take the values −1, 0, and 1 for a total of four states at $n = 2$. When $n = 3$, there are nine degenerate states. In general, the overall degeneracy is n^2. However, in any real atom there are interactions and subtleties that "break" this energy degeneracy.[1] These slight differences gives us some of the most accurate means to determine whether each refinement of our mathematical model is correct or not.

The actual shape of these orbitals are determined by two factors: (1) the radial probability density and (2) its angular dependence. The functions describing this angular pattern are called spherical harmonics. These mathematical functions show up in a variety of different contexts in physics: gravitational theory, electromagnetic radiation, even three-dimensional acoustic problems. These functions are what give the orbitals their characteristic spherical, tear-drop, or ring shapes.

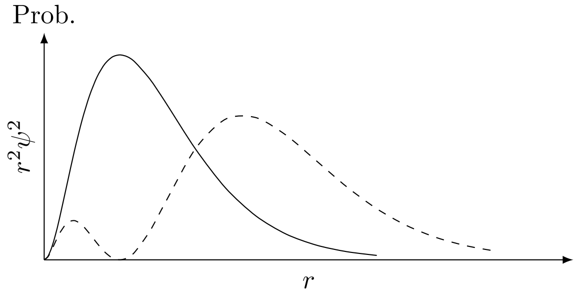


Figure 28.2: Probability of finding an electron at a particular radius in the atom ($n = 1$ solid, $n = 2$ dashed)

In general the radial function is an exponential decay, but because the volume decreases as r gets smaller, the actual probability of finding an electron at a particular radius looks like Figure 28.2. Technically there is a non-zero probability of finding the electron at any distance. The peak of this curve matches the Bohr radius given in equation (28.3).

So far in this explanation I've left out an important little tidbit called **quantum spin**. There are some similarities to the classical idea of a spinning top—it gives the electron a magnetic moment, for example—but overall it is better to think of this quantum spin as a uniquely quantum effect. For Schrödinger's equation we need to "bolt" on the idea of electron spin, but in quantum field theory the notion falls out quite naturally. In a way, an extra degree of freedom is expected since QFT is essentially a four-dimensional theory. For an electron, this quantum spin can only take the values of $\pm\hbar/2$.

We are starting to get enough information under our belt to begin to see the mechanical foundations of chemistry. Democritus in 400 BC was the first to advocate an atomic theory in which matter was composed of indivisible elements moving in random motion through the void. These elements were distinguished by shape and would lock together in accord with those shapes. Over two millennia later, we find out that he was essentially correct.

For example, the structure of the periodic table is a consequence of the Pauli exclusion principle. As protons are added to the nucleus, electrons are added to the exterior of the atom to balance the electrostatic charge. These electrons are added to the lowest energy levels first. Electrons with excess energy are said to be in an **excited state** and will eventually return to the **ground state** with the release of a photon of energy.

So we start with hydrogen with one electron in the 1s orbital.[2] Helium has another electron which can also sit in the 1s orbital if the spin of the two electrons are anti-parallel. This is denoted 1s^2. Lithium has three electrons. Since there is only a two-fold degeneracy in the $n = 1$ energy level, the third electron will sit in the 2s orbital. But the 2s orbital has an eight-fold degeneracy so the next eight elements gradually fill this shell until we get to neon.

I mentioned earlier that this energy degeneracy is broken in a real atom. In particular the shapes associated with different angular quantum numbers tend to

[1] The quantum number m is called magnetic because it's degeneracy can be broken with a magnetic field.

[2] There is a labeling convention that goes back to spectroscopic studies to label these orbitals by a leading number corresponding to the n value and a letter corresponding to the ℓ value. The letters are s, p, d, and f for $\ell = 0, 1, 2, 3$ respectively.

favor the more spherical shapes, or lower ℓ values. So, in general, the orbitals fill from s to p to d to f if possible given the value of the principle quantum number n. Thus, the orbital configuration for boron is $1s^2 2s^2 2p^1$.

After neon, things get a little more tricky. This is because the energy level of the 3s orbital is actually lower than the 2d orbital. In fact, the energy order of the orbitals is

$$\text{1s 2s 2p 3s 3p 4s 3d 4p 5s 4d 5p 6s 4f 5d 6p 7s 5f 6d 7p 8s 5g 6f 7d 8p}$$

Filling the electron shells according to this order will reproduce the period table (with a couple of exceptions).

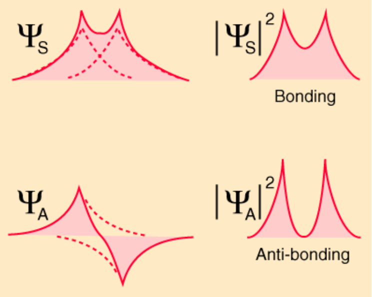


Figure 28.3: The electron fields from two hydrogen atoms can overlap constructively or destructively (image credit here)

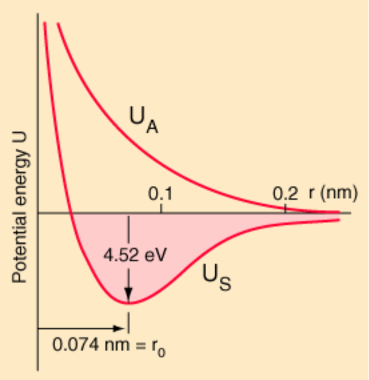


Figure 28.4: Energy levels for the two overlap patterns in Figure 28.3 as the atoms approach one another (image credit here)

The covalent bond is also a uniquely quantum mechanical effect. Consider the hydrogen molecule H_2. As the electron fields from the two atoms overlap, they may either constructively or destructively interfere (see Figure 28.3). It happens that the interference type is related to the quantum spin of each electron. If the two electron spins are aligned (parallel), the two waves constructively interfere and the overall energy of the system decreases. This reduction in energy is called the **exchange interaction**.

There is a point of equilibrium in which the electrostatic repulsion of the two electrons is compensated by this exchange interaction (see Figure 28.4). At this point the wave function for the electron pair is largest between the two hydrogen nuclei—we say the two are "sharing" the electrons. This constructive overlap is responsible for every covalent bond in chemistry.

A secondary consequence of the covalent bond is that it induces an electric dipole moment in the molecule by localizing the electrons between the nuclei. This effect is most pronounced when hydrogen is involved because it creates the maximum exposure of the positive nucleus. This is the source of the **hydrogen bond** which is essentially electrostatic in nature. The hydrogen bond is responsible for many of the properties of water (including the fact that ice is less dense than cold water), the structure of proteins, and holds the DNA molecule together.

The exchange interaction is also responsible for magnets. You remember from Lecture 24 that every spinning charge acts like a magnet. So, every atom is also magnetic because of both the intrinsic and orbital spin of all its parts. However, in classical mechanics the alignment of these atoms is completely random without any long-range order. Part of the reason for this is that magnetic fields do no work. This means that classical mechanics cannot explain the existence of bar magnets!

There are actually three forms of natural magnetism: ferromagnetism, paramagnetism, and diamagnetism. **Ferromagnetism** is the strongest type and is exhibited in the typical bar magnet. **Paramagnetism** is the type of magnetism involved in materials we would call "magnetic" in the sense that a natural magnet will "stick" to the substance. This is due to an induced dipole magnetic moment which acts to attract the magnet to the material. **Diamagnetism** occurs in "non-magnetic" materials like aluminum or plastic. This effect is actually repulsive and is ultimately due to Lenz's law (page **??**) applied on the atomic scale.[3]

But ferromagnetism has no classical analog. It is another manifestation of the exchange interaction which creates a lower energy state when the magnetic moments of the atoms are parallel. In a way we have a "magnetic covalent bond" which forms and yields an alignment of the atomic magnetic moments to create a macroscopic dipole moment.

Actually, ferromagnetism is our first step into solid state physics.[4] Classically we have already imagined a solid like set of balls connected by springs (cf. Lecture 13). We now hope to see what happens when we apply quantum mechanics to a collection of atoms locked together in a solid.

[3] This phenomena has been used to stabilize some interesting levitation experiments. See here to read about levitating live frogs...

[4] Solid state physics is a branch of condensed matter physics, which includes the study of all phases of matter: gas, liquid, and solid. It also studies the electric and magnetic properties of matter including more exotic phenomena like superconductivity. Nanotechnology (the manipulation of matter on an atomic scale) is also considered a branch of condensed matter physics.

Take another look at Figure 28.4. Notice how when the separation between the hydrogen atoms is large the electrons are both at −13.6 eV. As they approach one another this equality splits. If we add another atom to the system, the energy levels split again. Add a whole crystal of atoms and the energy levels form an **energy band** of closely spaced levels. This allows the electrons to gain a small amount of kinetic energy to move rather than being locked into one particular energy level.

Each orbital from the original atoms creates an energy band (see here for an interesting diagram). Frequently there are gaps between the bands—energy levels inaccessible to the electrons. As the atoms collect into the crystal, the electrons fill the lowest energy bands first. These bands are either completely full or completely empty. Bands that are filled are called **valence bands** and those that are empty are called **conduction bands**. It is possible that the bands overlap—if that occurs between the last valence band and the first conduction band, then the electrons are free to move and we have a conductor. On the other hand, an energy gap creates an insulator. See Figure 28.5.

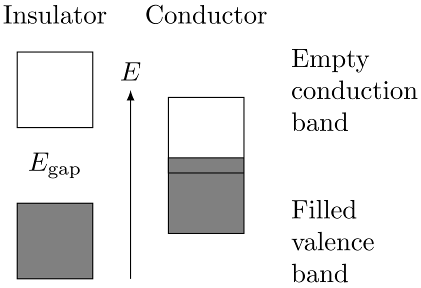


Figure 28.5: Energy band description of insulators and conductors

But this description ignores the effect of temperature on the energy distribution of the electrons. At any temperature, there is a probability that any individual electron will gain enough kinetic energy to overcome this energy gap. It can be shown that the density of electrons that do this is related to both temperature and the energy gap by

$$\rho = (AT^{3/2})\exp(-E_{\text{gap}}/2kT) \tag{28.5}$$

where $A = 0.00805\ \text{mol/m}^3$. The important factor is the exponential. Unless the energy gap is on the order of an electron volt, this temperature effect is completely negligible. For example, silicon has a band gap of 1.1 eV, so at 300 kelvin (approximately room temperature), equation (28.5) implies an electron density of 3.28×10^{-8} moles of electrons per cubic meter. Though small, this is enough to allow a small amount of current to flow.

So even if the bands don't overlap, there is a second way for a material to conduct electricity. Those materials with an energy gap around 1 eV are called **semiconductors**.

One of the reasons for the tremendous utility of semiconductors in electronics is the ability to precisely control the electrical properties through **doping**. If we diffuse a small amount of phosphorous (typically via phosphine gas PH_3) into pure silicon we introduce an extra energy level just below the conduction band filled with extra electrons because phosphorous has five outer electrons but the lattice structure of silicon only supports four. This is called an **n-type semiconductor** because of the addition of negative charges.

We can do the opposite too by diffusing boron (typically via diborane gas B_2H_6) which introduces an empty energy level just above the valence band. This gives the electrons trapped in the valence band a bit of room to move. If an electron jumps up to this extra energy level it leaves behind a "hole" in the valence band. A neighboring electron can jump into it leaving behind another hole. This hole acts exactly like a positively charged electron in that it moves in the opposite direction as the electrons with the same speed and effective mass. Though an artifact of the gaps in the electron "gas" produced by introducing boron, it is convenient to consider this hole as particle in its own right,[5] so we call this a **p-type semiconductor**.

But this is just the beginning. Now take p-type and n-type materials and put them together. This is called a **pn junction** and is fundamental to any integrated circuit. At the interface between the two materials, the free electrons in the n-type materials jump over to holes in the p-type material.[6] This creates a separation of charges which creates an internal voltage difference across the interface. This voltage difference opposes the flow of electrons and an equilibrium state is reached

[5] In fact, when Dirac was working out a relativistic version of Schrödinger's equation (27.2), he predicted the existence of the positron in the very same way.

[6] Notice that the p-type material has a net negative charge and the n-type material has a net positive charge—the naming convention gets a bit confusing at this point.

as in Figure 28.6. This internal voltage is dependent upon the nature of the two materials making the junction.

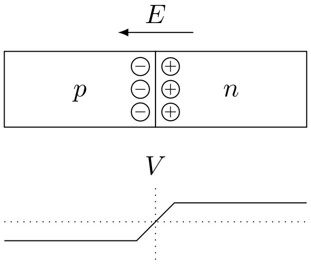


Figure 28.6: The equilibrium state of a *pn*-junction and the internal potential that must be overcome for current to flow freely.

Suppose we connect the positive terminal of a battery to the p side of the junction and the negative terminal to the n side. What happens is that the battery siphons off the electrons that collected in the p-type material (if the battery voltage is large enough to overcome the intrinsic voltage of the junction which is about 0.6 eV for a silicon-based diode). This disturbs the equilibrium and allows more electrons to flow from the n-type material and current flows freely.

Hook the battery up the other way and the current is blocked. This is because the battery voltage must exceed the internal voltage for current to flow. But as the current flows, it contributes to this internal voltage (like a capacitor) and it builds up until it is equal to the battery voltage.

So the pn junction acts like a one-way street for current. When we treat this pn junction as an electronic component it is called a **diode**.[7]

As current flows in a diode, each electron jumps from the conduction band in the n-type material to the valence band in the p-type material. When the diode is forward biased these levels are separated by an amount equal to the internal voltage at equilibrium. This means that each electron drops in energy just like when the electron in an exited atom falls to its ground state. However, not every pn junction will emit photons: the change in momentum has to be lined up too.[8] If so, the band gap is said to be "direct" and the energy drop is released as a photon (silicon based diodes have indirect band gaps). If the drop creates a photon with a frequency in the visible range, we have a **light-emitting diode**, or LED. Run this process in reverse and you have a **solar cell**.

We can't leave this subject without talking about the **transistor** which is essentially two diodes placed back-to-back. A *pnp* transistor has an internal voltage pattern that looks like a potential barrier to current. By controlling the doping levels we can control which side has higher and lower potential. The higher potential end is called the **collector** and the lower potential the **emitter**.

The central n region is called the **base** of the transistor. By applying a voltage to the base we can control the height of the potential barrier within the transistor. When the voltage on the base pulls the potential barrier down, current flows from the collector to the emitter (these names are a consequence of the fact that electrons move in the opposite direction because they have negative charge).

So a transistor acts as a miniature electronic switch which makes it ideal for computers. It also can be used as an amplifier since the base voltage—which uses a small amount of current—can control a large flow of current between the collector and emitter. Every integrated circuit is a combination of seemingly endless combinations of these diodes, transistors, and more.

Next week we continue our studies down in scale to talk about the nucleus of the atom. Each nucleus is composed of positively charged protons and particles without electric charge called neutrons. We will discuss how each nucleus is in tension balancing the electrostatic repulsion of its protons and the strong nuclear force. For larger nuclei this balance becomes unstable and natural radioactivity is a result. This tension also offers a way of unleashing the power of the strong nuclear force in controlled (and uncontrolled) ways.

[7] This terminology actually comes from the days of vacuum tubes—which are still used in some industrial applications (see here).

[8] This is true of the atom also: transitions between any two states is not possible in general. But typically the angular momentum states are degenerate so there is always a way to get from one energy level to another.

Lecture 29

Nuclear Energy

Read sections 31.1–31.7 and 32.1–32.5

Just a casual observation of the periodic table will uncover a problem: the atomic masses do not increase in step with the atomic number. One would expect that helium would be twice as massive as hydrogen, lithium three times, etc. But not so—its 1, 4, 7, 9, 11ish, 12, 14, 16, 19 for the first nine elements. And what is up with boron at 10.8?

Initially it was supposed that, for example, nitrogen had 14 protons in its nucleus (which accounts for the mass) that had somehow swallowed seven of its electrons (which accounts for the charge). However, no arrangement of 21 particles can correctly account for its quantum spin.

In 1920, Rutherford offered the idea of a third subatomic particle in the nucleus. Without charge and with mass similar to the proton, this **neutron** was later discovered in 1932. The atomic number of the atom describes the number of protons and drives the chemistry of the element (by attracting the same number of electrons), and the remaining mass is provided by the neutrons.

We can even explain our "11ish" mass for boron as the average mass of several **isotopes**, or atoms with the same number of protons but a different number of neutrons. Isotopes have different masses but are indistinguishable chemically. In the case of boron we have two stable isotopes: boron-10 and boron-11 with five and six neutrons respectively.[1] The relative abundance of the two are 20% and 80% which yields an average mass of 10.8.

The neutrons stabilize the nucleus by pitting the strong nuclear force against the electrostatic repulsion of the protons. Although protons do participate in the strong force, their repulsion is too great to allow them to bind together: helium must have at least one neutron.

On the other hand, the neutron needs the proton too. In isolation, the neutron is unstable and decays into a proton and an electron with a mean lifetime of about 15 minutes. But in the nucleus the neutron can be stable.[2]

However, when the nucleus becomes too large these balancing acts begin to fail. The heaviest stable isotope is lead-208. Nuclei heavier than this are **radioactive** meaning they decompose and release nuclear radiation.

A common way for nuclei to shed excess size is by alpha decay. One of the first forms of radiation studied, the α-particle was later determined to be a pair of protons and neutrons (an helium nucleus).[3] This alpha radiation is easily shielded by a piece of paper or even just a few centimeters of air.[4] This mode of decay is said to be **transmutation** since it actually changes the chemical element

[1] These are usually denoted $^{10}_{5}$B and $^{11}_{5}$B but it is so much easier to write, type and say boron-10 and boron-11 that I will almost exclusively do so.

[2] It may look like this validates the idea of the neutron as some sort of bound state between the proton and the electron. But the story is a more complicated than that. We will pick this up again in Lecture 30.

[3] This is the source of helium on our planet. Every helium balloon you see is filled with the byproduct of billions of years of radioactivity.

[4] Smoke detectors work by using an alpha radiation source. The alarm is designed to sound whenever the transmission of the α-particles is blocked (presumably by smoke).

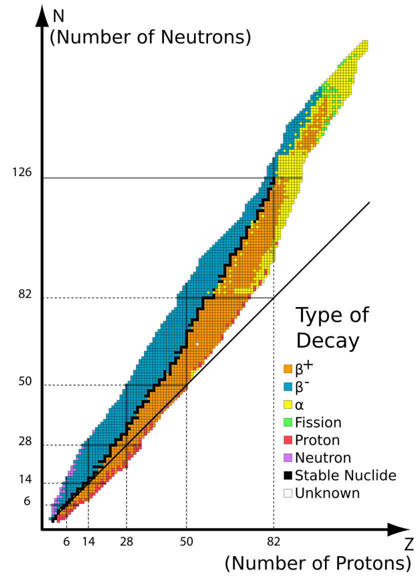


Figure 29.1: Normal decay type by isotope (image credit here)

by reducing the atomic number by two.

An imbalance between the number of protons and neutrons can also cause instability in a nucleus. If there are too many neutrons, it is typical for one to decay into a proton and eject an electron. This is called beta decay and is another transmutation process (adds one to the atomic number). These particles require more shielding—five millimeters of aluminum is typical.

If there are too many protons, one of them will transform into a neutron and eject a positron in the process (subtracting one from the atomic number). Based on the discussion from Lecture 27, we should not be surprised to see this time-reversed, anti-matter twin of the process of neutron decay. This is also called beta decay and the two particles are often labeled β^- and β^+ to distinguish them.

Alpha decay is the typical transmutation process for heavy nuclei while beta decay is typical for light nuclei. Other decay processes are possible, but the vast majority are one of these two. See Figure 29.1 for a complete picture.

It is not uncommon for these transmutation processes to leave the nucleus in an excited energy state. And like the atom, the nucleus will emit a photon of electromagnetic radiation to release this energy. These photons are called gamma rays and typically have millions of electron-volts of energy. The higher the energy, the more shielding is required to block this radiation. Three inches of lead is a generally accepted norm (any material will work including concrete and even dirt—you will just need more of it).

All three of these types of radiation are classified as **ionizing radiation** because they are energetic enough to ionize an atom (tens of electron-volts). As such they can do significant biological damage in sufficient quantity. The **radiation dose** is calculated as the amount of ionizing energy absorbed divided by the mass of the absorbing material. The SI unit of one joule per kilogram is called a gray.[5]

But not all radiation has the same effect on our physiology. Dose for dose, alpha particles and other nuclear fragments do 20 times more damage than beta particles and gamma rays. Since neutrons do not carry charge, they interact indirectly through sheer kinetic energy and their effect therefore varies. The most damaging neutrons are in the millions of electron-volts and do damage comparable to alpha particles. This variation of biological impact is summarized in a number called the **relative biological effectiveness**, or RBE.

The absorbed dose multiplied by the RBE gives us the **biologically equivalent dose** for the radiation exposure. Based on the gray, the SI unit is called the sievert.[6] The rule of thumb is that one sievert will cause nausea and over six is lethal. Our natural background radiation is approximately 2.4 millisieverts per year—about half from radon gas with the rest split evenly between cosmic radiation, terrestrial radiation, and our food and water. We are also exposed to an additional 0.4 millisieverts through various medical procedures (see here for details).

As was mentioned in Lecture 27, natural radiation is a quantum mechanical effect. The electrostatic repulsion of the protons increase as the distances decrease, but once we get within the size of the nucleus, the strong force overwhelms the electric force and pulls the whole thing together. So there is a potential energy barrier that holds the nucleus in place. Since the wave functions of the most energetic nuclear material extend beyond this barrier, there is a small probability that a particle will manifest outside. Thus the nuclear material is pushed away and radiation is the result.

So the actual emission of radiation is an event based on probability. This probability is independent of time, so at any one moment the percentage of nuclei actually radiating is fixed. This probability is called the **decay constant**, λ for the material. Thus, the rate of disintegrations[7] is given by

[5] The more traditional unit is the rad (radioactivity absorbed dose) which is one-hundredth of a gray, so 100 rad equals 1 gray.

[6] It is much more common (though strongly discouraged) to see this equivalent dose quoted in "rem" which is based on the rad.

[7] One disintegration per second is called a becquerel. Another common unit is disintegrations per minute or bpm.

$$\frac{\Delta N}{\Delta t} = -\lambda N \tag{29.1}$$

This quantity is called the **activity** of the material. From this it follows that the amount of substance left after a particular time period is

$$N = N_0 \exp(-\lambda t) \tag{29.2}$$

It can be shown from this equation that the mean life-time τ of a particle is equal to $1/\lambda$. The **half-life** of the material is the amount of time it takes for half of the initial substance to disintegrate and we see that

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2 \tag{29.3}$$

The half-life of carbon-14 is 5730 years, so its decay constant λ is 3.84×10^{-12}.

Radiocarbon dating works by comparing the amount of carbon-14 to carbon-12 in a substance. Natural carbon-14 is created by cosmic rays interacting with the nitrogen gas in the atmosphere. A dynamic equilibrium exists between the radioactive decay and this cosmic generation which fixes the amount of carbon-14 in the atmosphere. As such, living creatures breathe and consume both carbon-12 and carbon-14 in a fixed ratio.

When a plant or animal dies, the level of carbon-14 becomes depleted over time. By measuring the radioactive activity in a specimen we can use equation (29.1) to determine the amount of carbon-14 currently present. Measuring the amount of carbon-12 gives us a way to estimate the original amount of carbon-14. Using equation (29.2) we can get to an approximate time of death:

$$t = \frac{1}{\lambda} \ln\left(\frac{N_0}{N}\right) \tag{29.4}$$

where N is the current amount of carbon-14 inferred from the measured activity of the sample and N_0 is the original amount of carbon-14 inferred from the amount of carbon-12 in the sample.

So far in this lecture we have focused on natural forms of radioactivity and transmutation. We now turn to consider how we can use this information to induce various nuclear interactions. The most important item to consider in this context is the nuclear binding energy curve in Figure 29.2. This curve represents the net energy needed to break apart stable nuclei into their component parts.

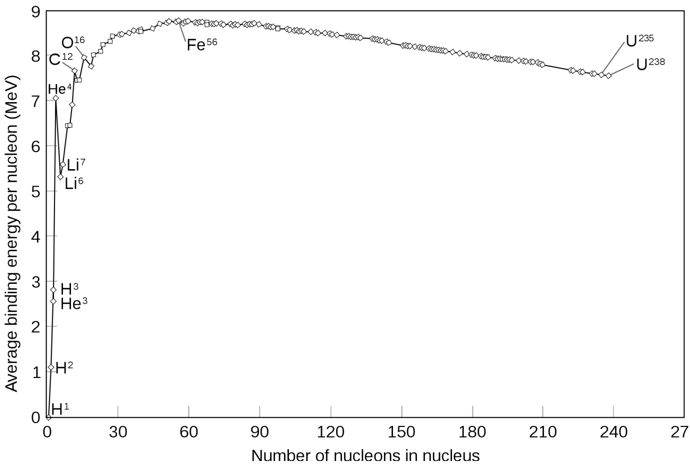


Figure 29.2: Nuclear binding energy curve (image credit here)

Note that this is not the energy required to overcome the strong nuclear force and break apart the nucleus (the “activation energy” to use the chemical term). If this barrier energy is overcome, electrostatic repulsion will create kinetic energy pushing apart the protons to infinity. The binding energy is the attractive energy from the strong force minus the repulsive energy from the electrostatic force and represents the average energy level of the nuclear material in the nucleus. Iron-56 is the most stable of all nuclei with a binding energy of 8.8 MeV per nucleon.[8] The high binding energy for helium-4 (at 7.1 MeV per nucleon) explains why heavy radioactive nuclei tend to decay using alpha particles rather than tritium or lithium-6 or some other combination of protons and neutrons.

Using Figure 29.2 we can calculate the energy involved in various nuclear reactions. **Nuclear fusion** occurs when we combine nuclei to generate a larger composite (we called this an inelastic collision in Lecture 8). The fact that the binding curve goes up initially indicates that fusion in exothermic: more energy is released when the end products combine than required to break apart the beginning nuclei. This is the kind of nuclear process that powers the sun.

The binding curve increases up to iron-56 after which fusion is endothermic: it requires a net expenditure of energy to create these nuclei. But this also implies that the reverse process, **nuclear fission**, can be used as an energy source. This is the kind of nuclear process used in nuclear power plants today.

In either nuclear fusion or fission, we can use the binding energy curve to calculate the net energy requirements for the process. However, there is an easier way using the so-called **mass defect**. This is a statement of the fact that the mass on the two sides of a nuclear reaction do not balance. The difference is simply a consequence of $E = mc^2$ and the binding energy involved. Every difference of one atomic mass unit corresponds to 931.5 MeV of energy. When one takes into account the "mass" of the binding energy, the equations balance.

For example, let's calculate the binding energy of the helium atom. We have two protons, two neutrons, and two electrons. The atomic masses of these particles are in Table 29.1. The total of all six particles is 4.032 980 atomic mass units while the helium atom is only 4.002 602. This is mass deficit of 0.030 378 atomic mass units, so the helium-4 atom has a binding energy of 28.3 MeV.

Electron	0.000 549 amu
Proton	1.007 276 amu
Neutron	1.008 665 amu

Table 29.1: Atomic masses of the three particles that make up the atom

The atomic masses of the isotopes are readily available (here for example, or look up the wikipedia article "isotopes of helium", etc.). In order to figure out the energy released in a particular nuclear reaction, we simply calculate the difference in mass and convert it to energy via $E = mc^2$.

The first man-made fission reaction involved bombarding uranium-235 with neutrons. One possible reaction (of many) is the following:

$$^{235}_{92}\mathrm{U} + ^{1}_{0}n \rightarrow ^{144}_{56}\mathrm{Ba} + ^{89}_{36}\mathrm{Kr} + 3^{1}_{0}n$$

The atomic masses for uranium-235, barium-144, and krypton-89 are 235.043 930, 143.922 953, and 88.917 636 respectively. The total mass on the left-hand side of this reaction is 236.052 590 and the right-hand side is 235.866 580, which leaves a mass defect of 0.186 010 atomic mass units. This means that each fission reaction generates 173 MeV of energy.

On average, the fission of uranium-235 yields 215 MeV of energy with 2.4 neutrons as a byproduct. These extra neutrons are typically too energetic to initiate a secondary reaction unless the concentration of uranium-235 (relative to uranium-238) is in excess of 90%—which is called weapons-grade uranium.[9] However, a moderator (like graphite) can be used to slow the neutrons down in order to create a controlled chain-reaction. Nuclear reactors work this way with a smaller concentration around 20% and harness the energy through a high-efficiency steam engine. The amount of energy generated by the fission of one kilogram of uranium-235 could easily provide the energy needs of one person for a lifetime.

The problem with nuclear fission as an energy source is the radioactive waste products. On the other end of the scale we have fusion which powers the sun. Fusion packs more power per kilogram of fuel which can be seen by comparing the slopes on the left and right of the binding energy curve in Figure 29.2. One

[8] A **nucleon** is a catch-all term to describe both protons and neutrons.

[9] Plutonium-239 will also support a chain reaction and can be generated by bombarding uranium-238 with neutrons.

of the most promising nuclear reactions for fusion on earth is the combination of deuterium (hydrogen-2, atomic mass of 2.014 102) and tritium (hydrogen-3, atomic mass of 3.016 049):

$$^{2}_{1}\mathrm{H} + ^{3}_{1}\mathrm{H} \rightarrow ^{4}_{2}\mathrm{He} + ^{1}_{0}n$$

The mass on the left-hand side of this reaction is 5.030 151 and the right-hand side is 5.011 267, for a mass deficit of 0.018 884 atomic mass units. This reaction produces 17.6 MeV of energy, or 3.52 MeV per nucleon. The fission of uranium-235 was about 0.915 MeV per nucleon, so pound-for-pound this reaction is almost four times more powerful—with helium as the main waste product.

There are two problems with this approach. The first is the fuel: deuterium is relatively plentiful and could be extracted from sea water, but tritium has a half-life of about 12 years (decay product is helium-3). So there is not much around—it would have to be generated (by bombarding lithium-6 with neutrons). This would take some work, but is possible.

The real problem is that no one has yet been able to control these fuels with enough precision to consistently overcome the "activation barrier" to release the fusion energy. This high level of energy requires the hydrogen isotopes to be in a plasma with a temperature over 450 million kelvin. These temperatures have been reached—the problem is constricting this plasma to create some sort of **magnetic confinement fusion**. Plasma is electromagnetically active which makes it extremely unstable and unwieldy. The tokamak uses a toroidal (donut) shaped magnetic field to perform the necessary confinement. Several experimental tokamaks have been constructed with some success and research is ongoing in order to produce a workable structure.

Another approach to the problem is called **inertial confinement fusion**. Essentially the idea is to fire a high-energy laser or electron beam at a small pellet of the deuterium/tritium fuel. The beam burns off the outer layer explosively which drives an implosion of the inner layer. If a critical density and temperature can be reached the fusion reaction will take place—just like a tiny atomic bomb.

Next week we finish off the course by investigating the very edge of what is known in physics. The decay of the neutron is our stepping off point to talk about the strong and weak nuclear interactions. We will find that protons and neutrons are made of quarks and we will discover we have overlooked the neutrino in our previous discussions. Quarks will help us make sense out of the "particle zoo" in high energy physics.

This is the "standard model" of elementary particles and our two nuclear interactions involve the exchange of new physical properties in the context of quantum field theory. If we have time, we may even be able to talk about why people think there is something beyond the standard model and whether string theory is that final answer.

Lecture 30

High Energy Physics

Read section 32.6–32.7 and review Lecture 27

Our humble neutron, which appeared as an add-on to the structure of the atom, turns out to play a central role in both the creation of nuclear fuel and the unleashing of nuclear energy. We will start this final lecture addressing one loose end from the previous one: neutron decay.

With regard to neutron decay, I left out an important fact from Lecture 29. When the neutron decays three particles are produced not two. The extra particle is called a **neutrino** because it has no electric charge and is nearly massless.[1]

We now seek an explanation in terms of quantum field theory for the decay of the neutron. We can do this by considering some key Feynman diagrams. We postulate the existence of a new boson to mediate the interaction that controls this decay. The reason for this is that each vertex in a Feynman diagram always forms a triad with two fermion lines and one boson. We cannot have a fermion decaying directly into other fermions. Since this is called the weak nuclear interaction, we call this new particle the **weak boson** (see Figure 30.1).

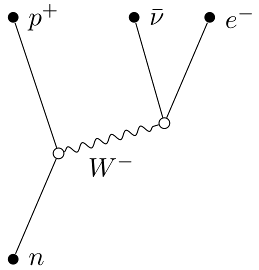


Figure 30.1: Feynman diagram for neutron decay. The weak boson creates valid vertexes for the interaction.

We know a few of things about this boson. First, it is short-range which means it is massive—about 80 GeV/c^2. It can be shown that equation (27.8) yields the potential formula

$$V \propto \frac{\exp(-r/a)}{r} \tag{30.1}$$

where $a = \hbar/mc$ represents the range of this potential. In QED the range is infinite because the mass of the photon is zero. Based on the mass of the W-boson, the range of its influence is on the order of 10^{-17} meters. Second, it decays with a very short time-frame—about 10^{-25} seconds which is a consequence of its large mass. In this case, the W-boson decays into an electron and an anti-neutrino. Finally, it is charged since it converts a neutral object (the neutron) into a charged one (the proton). This means it has an anti-particle twin with the opposite charge.

If Figure 30.1 were a diagram from QED, we would interpret the decay of the weak boson as a matter-antimatter creation event. This is why the neutrino is labeled as its anti-particle. But the new characteristic of this diagram is that it pairs different particles together: the neutron with the proton and the electron with the neutrino. There is more to this story, but we first need to develop some ideas on the strong nuclear interaction to pull it all together.

The fact that the neutron can decay seems to imply that there is still some substructure to be found. But we know from Lecture 27 that the creation of matter from pure energy cannot be ruled out. However, there is more direct evidence of substructure: the neutron has a magnetic moment.

[1] Originally, the neutrino was assumed to be massless. The phenomena of neutrino oscillation measured from the sun in 1998 proves that the neutrino must have some mass. The current upper bound is 0.3 eV, or about 1500 times lighter than the electron.

A magnetic moment is generated by a rotating charge (see Lecture 24). It is possible for an object with a net charge of zero to have a magnetic moment if the charge is distributed across the object (like a dipole). This seems to imply that there are little charged parts tied together to form the neutron. This is, in fact, the case—they are called **quarks**.[2]

The theory works if we assume the neutron is composed of three quarks: one "up" and two "down". Don't put too much thinking into the names: they are intended to merely distinguish the two as opposites.[3] These quarks have a charge of $+2/3$ and $-1/3$ respectively. If we combine two "up" quarks and one "down" quark we get the proton. And like the way electromagnetism is mediated by the exchange of photons, the strong interaction holds these triplets together by exchanging bosons called **gluons** (seriously...I'm not making this up).

The force that hold the nucleus together is actually the residual force left over from this strong quark interaction. Much like the way a magnet sticks to a refrigerator or a balloon sticks to the wall via static electricity, the fields from the bound quarks interact to hold the composite protons and neutrons together.

The Feynman diagrams for the exchange of gluons look just like those in QED, but to explain the three-way nature of the strong interaction requires some work: we need a three-valued "charge". This three-valued charge is called **colour**. Of course, the quarks are not actually red or green or blue, but there is some logic to this name. The basic rule of thumb is that quarks always combine to be colour-less or to create "white".

So the opposite of the red quark colour is green and blue together, and so on. You can imagine that the mathematics required to explain this is a bit complicated (this is why the SU(3) group is associated with colour). Figure 30.2 shows a schematic of what is happening. Remember, in quantum field theory you should imagine that all of these interactions are occurring simultaneously and continuously.

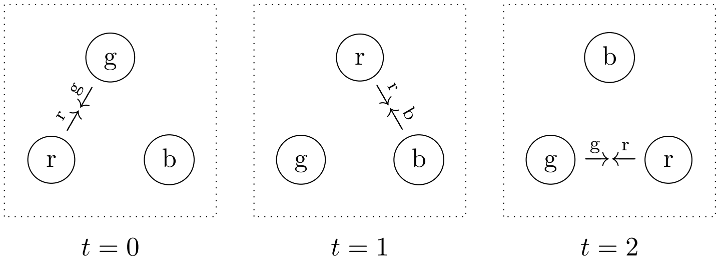


Figure 30.2: How the exchange of colour holds the proton and neutron together.

Let's look in a little more detail at the Feynman diagram for a single colour exchange. In Figure 30.3 the Q represents a quark of any kind and G is the gluon mediating the colour exchange. In this diagram, blue is on the left and red is on the right. The quark on the right emits the gluon which exchanges these colors. It takes red away and brings blue. But since the blue travels backward in time, we say the gluon is carrying red and anti-blue.

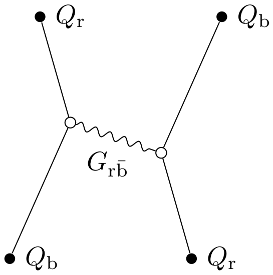


Figure 30.3: A fundamental diagram for the strong interaction showing the colour exchange between quarks.

The idea of "anti-color" brings the last piece of the puzzle into play: anti-quarks. With this same construct we can bind a quark and an anti-quark together. In this case, the anti-quark actually carries anti-colour, so they can combine in a colour-neutral way by being red and anti-red, for example.

These two ways of combining quarks explain the so-called **particle zoo**. In the decades before the theory of quarks was developed, many high energy collisions experiment were performed. Frequently new particles were formed in the collision of electrons or protons. Over time, this continual discovery of new particles stirred up quite a bit of discomfort among physicists. Around 1930, it seemed that the world could be explained with only three fundamental particles—now there were over 70 "fundamental" particles. Clearly something needed some explaining!

[2] If this seemingly endless substructure seems silly to you, you are not alone. When first proposed in the late 1960s, physicists were reluctant to accept the quark theory. But when predicted particles were found in the following decade (1974 to be precise), the theory carried the day.

[3] This is a common theme in particle physics: an appalling lack of imagination in naming things. We've already seen this in the "strong" and "weak" nuclear forces.

The current list of particles can be accounted for through the existence of six quarks. The triplet bound states are called **hadrons** and the doublet bound states are called **mesons**. Dozens of combinations have yet to be discovered, but all known particles are combinations of the fundamental six quarks in Table 30.4.

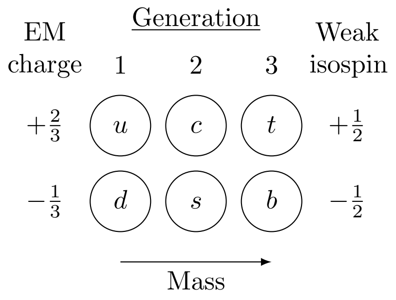


Figure 30.4: The six fundamental quarks: up, down, charm, strange, top and bottom.

The way this table is arranged is not random. The more massive ones are on the right and will decay through the weak interaction into the the less massive ones on the left. The quarks in the top row are said to have a weak isospin of $+1/2$ while those in the bottom row are $-1/2$. Each vertical pair is called a "generation" and the most likely decay path is within these generations (i.e., top to bottom, charm to strange, and down to up). It's also possible for the decay to skip to the next generation and even less likely to skip two. The one thing that cannot happen is a direct decay across the table. These facts are nicely summarized in Figure 30.5.

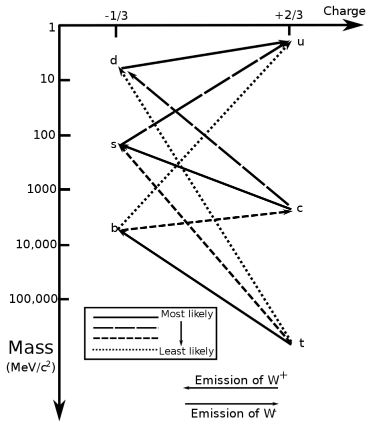


Figure 30.5: Possible quark decay paths via the weak interaction (image credit here)

So we come full circle. The quark composition of the neutron is *udd* while the proton is *uud*. So, when the neutron decays, this is really the decay of the internal d quark.[4] While manifesting the u quark, the weak boson carries away a full unit of negative electric charge and a full unit of negative weak isospin. Shortly afterward, the weak boson is mostly likely to decay using its least massive vertex: the electron/anti-neutrino pair.

Once again, it looks like we've got a theory that explains the facts. But nature is still not done. We've left out a particle called the **muon**. For all intents and purposes, the muon is nothing but a very massive electron (about 200 times greater). The redundancy is so striking that when it was first discovered (in 1937) Rabi famously joked, "Who ordered that?"

The muon has its own neutrino and in 1975 a third twin was discovered with its own neutrino. The new particle is called the **tauon** and is nearly 3500 times the mass of the electron. This particle is so massive that it can actually decay into hadrons (protons and neutrons), though it is more likely to decay into a muon or an electron.

These three pairs are called **leptons** and fall into three generations not unlike the quarks. In parallel with Figure 30.4 we have our six leptons in Figure 30.6. They decay through the weak interaction in a way that is similar to the quarks, but do not participate in the strong interaction at all.

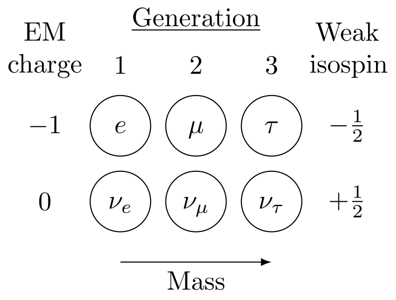


Figure 30.6: The six fundamental leptons: electron, muon, tauon each with a neutrino.

So let's summarize. The world is divided into two groups: bosons and fermions. Bosons mediate the three fundamental interactions. We have the photon, the weak bosons, and the gluons. The fermions form matter and are also divided into two groups: quarks and leptons. There are three pairs of quarks which participate in all three interactions. These quarks combine in doublets and triplets called mesons and hadrons. Ultimately, the proton is the only stable quark combination though the neutron is pretty close. The leptons also come in three pairs. All participate in the weak interaction, the electrons-like particles participate in the electromagnetic interaction and none participate in the strong interaction. Although the quarks can decay across generations through the weak interaction, the leptons do not (although the neutrinos do oscillate between the generations).

Of the three, the weak interaction is the most complicated. Every particle participates but in different ways. There is no single rule to how it works, although a pattern is there. And on top of it all, theory predicts the weak bosons to be massless. I couldn't break this news to you earlier. Technically, the theory of the weak interaction is DOA.

In 1967 a solution was discovered that involves two independent concepts. The first is the unification of the electromagnetic and weak interactions into a single one called the **electroweak interaction**. In this electroweak theory we have four bosons (W^+, W^-, W^0, and B^0) all of which are initially massless. Not much help so far.

[4] You may wonder why then doesn't the proton decay into a uuu combination? The answer is that the way the quantum spin of the quarks interact with the strong interaction requires this symmetric state to align all three spin states. This is possible (named the Ξ hadron), but the energy involved is higher than the proton. This makes the proton the least massive of all the hadrons.

The second concept is called the **Higgs mechanism** which postulates the existence of yet another entity called the Higgs field. The idea runs like this. At very high temperatures the four electroweak boson and the Higgs field exist in a kind of equilibrium. But when the energy level of the temperature drops below the mass of the weak bosons ($kT = 80$ GeV/c^2 or $T = 10^{15}$ kelvin), this equilibrium go unstable and the quantum symmetries "break"[5] and the Higgs field "mixes" with the electroweak bosons.

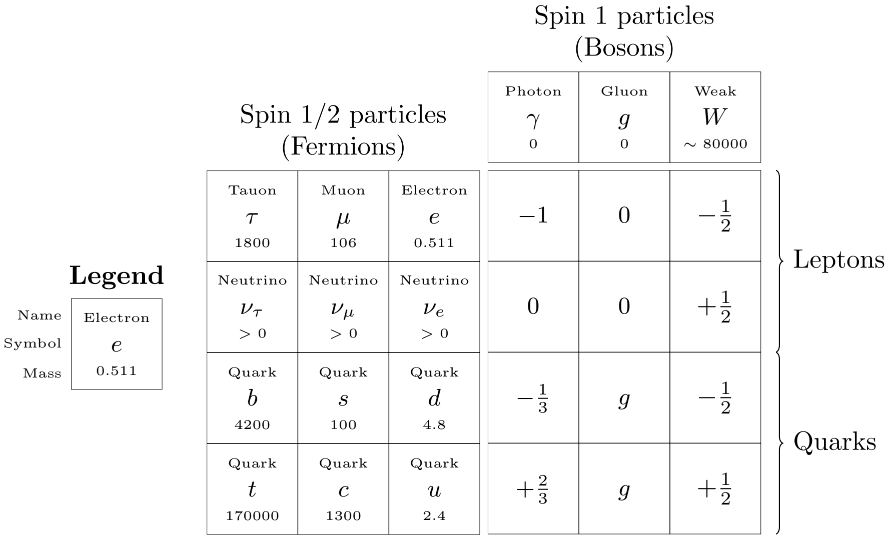
This mixing causes the W^+ and W^- bosons to acquire mass and the B^0 boson becomes the photon. We also have two other residual particles: the Z boson which is what is left of the original W^0 boson and the Higgs boson which is what is left of the original Higgs field. The Z boson has been observed (about 92 GeV/c^2) and acts like a heavy photon (it doesn't carry electric charge or weak isospin, so it doesn't change the nature of the particle). The Higgs boson has yet to be observed—this is one of the main objectives of the Large Hadron Collider.[6]

This approach may look convoluted and unmotivated. It is inspired by the BCS theory of superconductivity. Electrons pair up in a superconductor and act like bosons which gives the superconductor its non-classical behavior. A superconductor is a system involving massive "bosons" about which we know quite a bit. So it makes sense that this theory might act as a template to understand the weak boson.

Table 30.1: The observed masses of the weak bosons match well with those predicted by the electroweak theory (taken from here)

Boson	Measured	Predicted
W	80.398 ± 0.025	80.390 ± 0.018
Z	91.1876 ± 0.0021	91.1874 ± 0.0021

Nonetheless, I don't think anyone would call the theory "pretty". But it does work, and from a bottom-line perspective that's enough (see Table 30.1). The physicists working this out in the decades prior to 1980 or so seem to have expected this complexity to disappear in some overarching "theory of everything". As such, the combination of the electroweak theory and the strong interaction now goes by the unimpressive name of the **Standard Model** of particle physics. Figure 30.7 shows our new periodic chart for elementary particles.



Spin 1/2 particles (Fermions)			Spin 1 particles (Bosons)			
			Photon γ 0	Gluon g 0	Weak W $\sim$ 80000	
Tauon τ 1800	Muon μ 106	Electron e 0.511	-1	0	$-\frac{1}{2}$	Leptons
Neutrino ν_τ > 0	Neutrino ν_μ > 0	Neutrino ν_e > 0	0	0	$+\frac{1}{2}$	
Quark b 4200	Quark s 100	Quark d 4.8	$-\frac{1}{3}$	g	$-\frac{1}{2}$	Quarks
Quark t 170000	Quark c 1300	Quark u 2.4	$+\frac{2}{3}$	g	$+\frac{1}{2}$	

Legend

Name	Electron
Symbol	e
Mass	0.511

Figure 30.7: The elementary particles according to the Standard Model. (Inspired by a similar chart in Feynman's book *QED: The Strange Theory of Light and Matter*.)

The Standard Model is an extremely successful theory: one that has yet to make a prediction that is incompatible with experiment. Yet physicists continue to talk about the "next" theory. Why?

[5] This spontaneous symmetry breaking can be understood by example: consider what happens when a liquid freezes. As a liquid the molecules are oriented in all directions with no preference. But as the temperature drops, the molecules line up in a crystal structure and breaks the symmetry. Now, it is possible for a liquid to go into an unstable supercooled state. Some sort of asymmetric trigger is required to break the symmetry. But the point is that the preferred direction is not inherent in the physical laws governing the material. When the system moves to an unstable region (through lower temperature) the smallest trigger will lock the whole system in place.

[6] There may be ways to break the electroweak symmetry without the Higgs boson. See here for details.

The short answer is that the model appears incomplete. Clearly we need to incorporate gravity yet it is not clear how this can be done. The Higgs particle has yet to be found—what if it's not there? Also, there are things that happen in nature that the original model does not anticipate like parity violation[7] and neutrino oscillation. These features can be "bolted" on to the model to match experiment by tweaking and fine tuning the formulas, but a complete theory really ought to predict these features naturally. And then there are other things that you would expect the theory to explain that it does not. For example, is there a pattern to the particle masses?[8] How did the weak symmetry break and why did it fall out in this particular way? Why are there three generations of particles—are there more?[9] Finally there are suspicious coincidences: Why three generations of both quarks and leptons? Why is the charge of the electron and proton exactly the same?

It's these last two questions that have motivated physicists to postulate some sort of "supersymmetry" between the leptons and fermions which unifies the strong and electroweak interactions. Inspired by the success of the electroweak theory and even the memory of the unification of the electric and magnetic forces these models are called **grand unified theories** (GUT). None work. The main problem that they all have is that they predict the decay of the proton: just as the weak interaction decays the quarks and leptons down the three generations into less massive particles, so should this "supersymmetry" decay the proton down into electrons and neutrinos. No one has ever witnessed the decay of the proton—the current lower bound on the half-life of the proton is 10^{33} years and growing every day.

And then there is gravity.... This is where **string theory** enters the picture. But before I explain these speculations let's be very clear: we are deep in the wilderness here and the forest is very dark.[10] We are still waiting to see if we understand the electroweak unification correctly and we know that unifying the strong interaction has fundamental problems. The "Hail Mary" approach of string theory is a long shot indeed.

Nonetheless, the approach has prestigious roots: Einstein spent the last decades of his life working on it. He was inspired by the quirky observation that if one expands general relativity to five dimensions, Maxwell's equations pop out. There are problems, of course. There are no fermions, there are unobserved extra bosons, and it doesn't say anything about the strong and weak interactions. Einstein appears to have hoped to see an explanation for quantum mechanics come out of the theory, but it never did work.

But the "extra-dimension" seed had been planted and fifty years later string theory blossomed. Many theoretical facts about quantum strings vibrating in 10 dimensions were uncovered which yield supersymmetry with a new boson to mediate the force of gravity. The mathematics is obscure but in the end there appears to be five different options for a viable string theory. In the mid-1990's a single 11-dimensional version based on vibrating membranes was developed which incorporates the five as special cases. Currently much speculation (and controversy) is focused on this "M-theory" as the final unification of fundamental physics.

That's it. In Lecture 1 we began without even knowing how to define physics. Now we are at the very edge of what we know (and a bit beyond). It's been quite a ride and I hope you learned a bit in the process. Study hard and fare well.

[7] This parity violation implies that nature makes a slight distinction between matter and anti-matter. This could explain the apparent imbalance in the universe between the two.

[8] Note that the periodic table does offer an explanation for each atomic mass.

[9] Neutrino oscillation provides evidence that there are only three kinds of neutrinos, but it does not explain why.

[10] It is a bit of a pet peeve of mine that the hypothetical fringe of physics is often represented as fact. For example, there is no evidence for a "graviton"—it's just a guess that a quantum theory of gravity might have them. The unification of forces going back to the Big Bang is another: the "inflation phase" of the early universe is a way of solving the horizon problem, but this presupposes unification rather than providing evidence for it.

Index

aberration, 118
absolute zero, 88
acceleration, 16
action, 42
activity, 171
adiabatic, 100
alternating current, 144
ampere, 137
Ampere's law, 141
amplitude, 69
angular displacement, 50
angular frequency, 71
angular momentum, 60
anti-nodes, 111
aphelion, 63
arc-length, 49
Archimedes' principle, 82
atomic hypothesis, 93
atomic mass units, 94
attenuation, 124
Atwood machine, 29
Avagadro's number, 93
average velocity, 15

back EMF, 145
Balmer lines, 163
bandwidth, 148
base, 168
beat frequency, 113
beats, 112
Bernoulli's equation, 84
binomial theorem, 8
biologically equivalent dose, 170
birefringence, 124
black hole, 68, 154
blackbody, 97
Bohr model, 129
Bohr radius, 164
Boltzmann constant, 94
bosons, 130, 161
Brownian motion, 93
bulk modulus, 76
buoyant force, 82
butterfly effect, 85

capacitance, 135
capacitor, 133
Carnot cycle, 101
Celsius scale, 88
center of gravity, 58
center of mass, 47
center of mass frame, 47
central force, 35
centrifugal force, 31, 52
centripetal force, 34
coefficient of friction, 24
coefficient of restitution, 45
coefficient of thermal expansion, 88
coherence, 124
collector, 168
colour, 176
completely inelastic collision, 45
conduction bands, 167
conductivity, 138
conductors, 131
conservation of energy, 39
constitutive equation, 24
constraint forces, 23
constructive interference, 109
continuity equation, 83
continuum mechanics, 77
Coriolis force, 52
Coulomb's law, 131
critical angle, 117
critically damped, 73
cross product, 12
curl, 134
current, 44, 137

dark matter, 36
de Broglie wavelength, 157
decay constant, 170
decibels, 108
degenerate states, 164
degrees of freedom, 49
delta-v, 46
derivative, 7
destructive interference, 109
desynchronization, 151
deterministic chaos, 85
diamagnetism, 142, 166
dielectric, 135
dielectric constant, 135
diffraction, 122
diffraction grating, 123
diffusion, 96
diffusion constant, 96
diode, 168
diopters, 118
dipole moment, 132

dispersion, 107
displacement current, 149
divergence, 134
doping, 167
dot product, 12
double slit, 121
dynamic balancing, 61
dynamic equilibrium, 22
dynamic pressure, 84

eccentricity, 64
edge effects, 133
effective potential, 64
efficiency, 23, 38, 101
elastic collision, 45
electric constant, 133
electric field, 132
electric generator, 144
electrical interference, 145
electromagnet, 142
electroweak interaction, 177
elevation head, 84
ELI the ICE man, 147
EMF, 137
emissivity, 89
emitter, 168
energy, 38
energy band, 167
entropy, 102
equilibrium, 21
equipartition principle, 96
equipotential, 134
equivalence principle, 31, 154
ergosphere, 80
ether, 125
Euler-Lagrange equations, 53
event horizon, 80, 154
exchange interaction, 166
excited state, 165

Fahrenheit scale, 88
far-field, 135
Faraday's law, 143
fermions, 130, 161
ferromagnetism, 142, 166
Feynman diagram, 47
Fick's law, 96
field potential, 134
first law of thermodynamics, 99
fixed rotation, 49
flux, 132
focal point, 116
forbidden regions, 41
force couple, 55
four-momentum, 153
four-vectors, 13
fourier analysis, 113
free rotation, 49
free-body diagram, 23
frequency, 69
fundamental frequency, 111

gauge pressure, 82
Gauss' law, 133
general relativity, 154
generalized coordinates, 52
geometric optics, 116
geostationary orbit, 36
gluons, 176
grand unified theories, 179
gravitational sling-shot, 47
gravitational time dilation, 154
ground state, 165
gyroscopic precession, 61

hadrons, 177
half-life, 171
hard boundary, 110
harmonics, 111
heat, 40
heat capacity, 89
heat conduction, 89
heat convection, 89
heat engine, 100
heat radiation, 89
heat reservoir, 100
Heisenberg uncertainty principle, 158
hertz, 69
Higgs mechanism, 178
Hooke's Law, 69
hydraulic head, 84
hydrogen bond, 166
hydrostatic equation, 82

ideal fluid, 83
ideal gas, 94
ideal gas law, 94
ideal pulley, 23
ideal spring, 69
ideal string, 23
impedance, 106, 147
impetus, 20
impulse, 43
in phase, 106
in series, 138
incident wave, 110
incompressible, 83
index notation, 78
index of refraction, 117
inductor, 145
inertia, 27
inertial confinement fusion, 173
inertial forces, 31
inertial frame, 28
instantaneous velocity, 16
insulators, 131
intensity, 108
intensity level, 108

interferometer, 122
internal energy, 95
inviscid flow, 83
ionizing radiation, 170
irreversibility, 101
isobaric, 100
isochoric, 100
isothermal, 100
isotopes, 169

joule, 37

Kelvin scale, 88
Kepler's three laws, 35
kilogram, 22
kinetic energy, 38
kinetic equilibrium, 22
kinetic friction, 24
Kirchoff's laws, 138
Klein-Gordon, 159

lab frame, 45
Lagrangian, 42
Lagrangian mechanics, 52
laminar flow, 83
Larmor's formula, 150
laser, 161
latent heat, 90
law of gravity, 35
leakage, 133
length contraction, 151
lens, 116
lens equation, 116
Lenz' law, 143
leptons, 177
lever arm, 55
light clock, 150
light ray, 115
light-emitting diode, 168
line of action, 55
long-range, 24
longitudinal wave, 105
Lorentz factor, 152
Lorentz force, 139
Lorentz transformation, 151

magnetic confinement fusion, 173
magnetic constant, 139
magnetic poles, 139
magnetoelectric induction, 149
Malus' law, 125
mass, 27
mass defect, 172
mass flow rate, 83
matter wave, 157
Maxwell's equations, 149
Maxwell-Boltzmann distribution, 95
mechanical advantage, 23
mesons, 177
meter, 4
mode, 112
mole, 93
molecular mass, 94
moment of inertia, 59
momentum, 43
motional EMF, 143
multipole, 107
muon, 177
mutual inductance, 144

n-type semiconductor, 167
natural frequency, 71
Navier-Stokes equations, 84
neutrino, 175
neutron, 169
newton, 22
Newton's first law, 27
Newton's second law, 28
Newton's third law, 31
Newtonian fluid, 84
nodal line, 112
nodes, 111
non-conservative force, 40
normal forces, 23
normal stress, 75
nuclear fission, 172
nuclear fusion, 172
nucleon, 172
nutation, 61

Ohm's law, 138
optical power, 118
orbital precession, 67
osmosis, 96
out of phase, 106

p-type semiconductor, 167
parallelogram law, 9
paramagnetism, 142, 166
particle, 15
particle zoo, 176
pascal, 82
Pascal's principle, 81
passive components, 146
Pauli exclusion principle, 161
perihelion, 63
period, 33, 69
phase shift, 71
phase transition, 89
phasors, 145
photoelectric effect, 128
photon, 129
plasma, 91
pn junction, 167
Poiseuille's law, 84
polar coordinates, 35
polarization, 105
potential energy, 38

power, 38
pressure, 21
pressure head, 84
pressure stress, 76
principal axes, 60
principal quantum number, 164
principle of least action, 42
principle of least time, 118
principle of relativity, 27
propagation vector, 115
proportionality constant, 6

quality factor, 148
quantum electrodynamics, 159
quantum field theory, 47, 159
quantum spin, 165
quantum tunneling, 157
quarks, 176

radian, 50
radiation, 105
radiation dose, 170
radioactive, 169
radius of curvature, 33
range equation, 19
Rayleigh criterion, 123
reactance, 146
real image, 116
reference frame, 15
reflected wave, 110
reflection, 110
refraction, 115
relative biological effectiveness, 170
relativistic energy, 152
relativistic mass, 152
relativistic momentum, 152
resistivity, 138
resistor, 137
resonance, 73
rest energy, 153
rest mass, 152
right hand rule, 139
rigid object, 49
root-mean-square, 96, 144

scalar multiplication, 10
Schroedinger wave equation, 158
Schwarzschild metric, 154
Schwarzschild radius, 154
second, 4
second law of thermodynamics, 101, 103
self-inductance, 145
semiconductors, 167
shear modulus, 76
shear stress, 75
simple harmonic motion, 70
simple harmonic oscillator, 71
simple machines, 37
Snell's law, 117
soft boundary, 111
solar cell, 168
solenoid, 142
space-time interval, 152
spring constant, 69
stable equilibrium, 22
standard deviation, 96
Standard Model, 130, 178
standing wave, 111
static equilibrium, 22
static friction, 24
statistical mechanics, 95
steady state, 73
Stefan-Boltzmann constant, 89
Stokes' law, 84
strain, 21, 75
strange attractors, 85
stress, 21, 75
string theory, 179
superconductivity, 91, 161
superfluid, 161
support force, 23

tauon, 177
tensor, 13
terminal velocity, 18
thermal conductivity, 89
thermal expansion, 88
thermometric property, 87
thin-film interference, 124
third law of thermodynamics, 103
threshold of hearing, 108
time constant, 139
time dilation, 151
torque, 21, 55
transformer, 145
transient state, 73
transistor, 168
transmitted wave, 110
transmutation, 169
transverse wave, 105
triple point, 89
turbulent flow, 83
turning points, 41
turns ratio, 145

ultraviolet catastrophe, 97, 128
uniform circular motion, 33
unstable equilibrium, 22

valence bands, 167
Van der Walls equation, 95
vector addition, 9
vector basis, 10
vector components, 10
vector product, 12
velocity, 15
velocity head, 84

Venturi effect, 84
vibration, 69
virtual image, 117
virtual photon, 159
virtual work, 38
vis viva, 43
voltage, 134
volume flow rate, 83

watt, 38
wave equation, 106
wave front, 115
wave interference, 109
wave medium, 105
wave number, 107
wavelength, 107
weak boson, 175
weber, 139
weight, 22
Wien's Law, 128
work, 37
worm hole, 80

Young's modulus, 76

zero-point energy, 161