

Physics for the Life Sciences I

Version 0.8, January 3, 2013

Author:

Timothy McKay

University of Michigan

Table of Contents:

1. Physics and life	2
2. Standing up and staying still	33
3. Forces and structures	46
4. Understanding staying put	67
5. Getting around: friction and motion	93
6. Quantifying motion in one dimension	116
7. Getting started and moving around	134
8. Turning the corner: motion in 2 and 3 dimensions	155
9. What's happening: work and kinetic energy	202
10. What could happen: potential energy	222
11. Modeling isolated interactions: collisions	239
12. Mixing it up: oscillations	257
13. Energy at the atomic level	281
14. Keeping cool and staying warm	297
15. Using chaos: diffusion and life	307
16. Structures and processes in a random world	327
17. Floating: fluid statics including surfaces	343
18. Flowing: fluid dynamics including viscosity	377

1. Physics and life

- 1) Introduction to the course: life is physical
- 2) Three examples to set the stage
 - i. Modeling and the scaling laws
 - ii. Physical constraints and convergent evolution
 - iii. Diffusion and the sizes of things
- 3) Differences among the sciences and how you study them
- 4) Tools for this course
 - i. Volume and surface area
 - ii. Specific and total quantities: density and mass
 - iii. Trigonometry and essentials of calculus
 - iv. Orders of magnitude and scientific notation
 - v. Estimation
 - vi. Units of time, space, and mass; questions of scale
 - vii. A universe made of atoms
- 5) The natures of things: scalars and vectors
 - i. An example: the displacement vector
 - ii. Vector addition
 - iii. Rescaling: multiplying vectors by scalars
 - iv. Vector subtraction
 - v. Component notation: addition, subtraction, and equality
 - vi. Choosing a thoughtful coordinate system
- 6) Decoupled motion and vector components: relative velocity
- 7) Multiplying vectors by vectors: the scalar and vector products
- 8) Life's media: air and water

Physics for the Life Sciences: Chapter 1

1.0 Physics and Life

This book is intended for those who would like to understand life. It is especially well suited for those who aspire to one day add to our knowledge of life; for researchers in the life sciences.

One of the great achievements of 20th century science was to begin revealing the fundamental mechanisms of life. As these revelations emerged, the life sciences expanded from their traditional domains in biology to include chemistry, physics, mathematics, and engineering. The purpose of this book is to help you explore some very foundational elements of physics, especially those most important for understanding life. In it, you will learn some of the laws of physics and discover how they enable everything that life does. You will also see how the boundaries of biodiversity are defined by the constraints physical laws place on life.

This text is meant to support a year-long exploration of physics as it pertains to life. During this first half, we will focus on mechanical and thermal aspects of life, including the fluids in which we live and of which we are largely made. Here are some of the many questions we will address.

- How does the inexorable pull of gravity affect the sizes and shapes of organisms? What must they do to move around?

- How does inanimate matter apply forces? How can we predict what these forces will be?
- How do organisms use membranes, muscle, tendon, and bone to support themselves, get up and move about?
- How do people walk, birds fly, and squid swim?
- Why does it take so much effort to jog along at a constant speed? Why haven't organisms evolved wheels to make this easier?
- What is energy, and how do organisms take energy from or give energy to an object? What forms can energy take?
- What is temperature and thermal energy? How does life harness purely random thermal motion to get things done?
- How do organisms manage to survive winters and live in cold oceans?
- How do living things get along within life's media: air and water?

In the second half of the course, we will learn about several new aspects of the physics important for life:

- How do electrical forces give matter its strength? How does a watery environment enable life to build, manipulate, and break up large molecules freely?
- How does life send signals within an organism? Electric fields and potentials, electric currents and circuits, electricity and magnetism, the function of nerve cells.
- How do living things sense the world around them? Sound and light, imaging and detection.
- How can we extend our senses? Instrumental imaging.
- What is life built of? The elements, nuclei, radiation, and the origins of these.
- What conditions does life require, where do we find these in the universe, and why is finding life beyond what we see here on Earth so challenging?

Introduction to the course

Science is a great human endeavor. For centuries, millions of people have worked together to expand our collective understanding of what things make up the world and how those things interact. To explain, scientists seek consistent patterns in what they find and how things work; laws of nature. These laws are both proscriptive and restrictive. They tell us what can happen, but they also provide constraints on the possible. All the laws of science are provisional. By the time we call them laws they have been widely tested, but each remains open to revision.

These days, more than 350 years after Galileo's death, science has become a broad, complex endeavor. People speak of "the sciences", and include among them subjects like biology, chemistry, geology, medicine, astronomy, and physics. The long list gives an exaggerated impression of segregation among these fields. It is true that the focus of study, and often the methods adopted, vary from field to field. But much more unites the sciences than separates them. All rely on the same basic notion; that everything comes about as the consequence of a limited set of laws which govern what can happen. These are not laws of physics, or of chemistry, or of biology. They are not principles which you study in one class just for that class, or just to please your professor. They are laws of nature: unchanging, universal, and inescapable.

The particular focus in this course will be "physics". In the past, this word had the broadest meaning; it meant simply "natural science". The subject was everything in nature. At the time, most people believed there was much more to the world than nature. This *super*-natural world was the topic of "meta-physics". From a very early time, astronomy was considered separate. The precision and permanence observed in

the stars and planets seemed completely unearthly. This strange perfection suggested an utterly different reality, and astronomy was held apart from earthly science. Newton was the first to suspect that the heavens and the Earth might all obey the same set of laws. The last century of research has confirmed his suspicion to an astonishing degree, and now we speak of a united “astrophysics”.

As time went on and the successes of science grew, other disciplines were defined, carving off corners of the natural world for specialization. Chemistry first emerged around 1700, focusing on the elements of matter, their states, and how they come together to form the compound substances. Still more recent is biology; a term invented in 1802 by Gottfried Reinhold Treviranus, a German naturalist. Biology is defined today as “the division of physical science which deals with organized beings or animals and plants, their morphology, physiology, origin, and distribution”¹. During the 20th century some of the divisions which had grown up between the sciences began to dissolve. The connective fields of biochemistry, biophysics, physical biology, physical chemistry, chemical physics, mathematical biology, and biostatistics are all pursued in academic departments with their own professional organizations and dedicated journals.

Introductory science teaching has been slow to respond to this enhanced connection, often retaining a narrow disciplinary focus. This text is an effort to catch up; to reveal some of the exciting connections among the disciplines so important for the modern life sciences.

It is worth noting that biology is defined as a division of the physical sciences, as if to emphasize that there is no fundamental distinction between physical and life sciences. In this course, we will treat the suggestion that biology is “a division of physical science” in a deadly serious way. The most important idea in this course is that life is the outcome of physical laws, *and nothing else*. There is no “spark of life” which separates the animate and inanimate. Biology is not a subject apart from chemistry and physics, but rather the most complex and interesting application of them.

This assertion may surprise you. It may conflict with beliefs you have long held, with what you have been taught in the past, or with your gut feeling about things. This is good. Intellectual dissonance is the surest sign that you’re in a position to really learn.

I’m not going to insist, or even ask, that you *believe* this assertion. Instead we will take it as a starting point, a possibility worth thinking about and testing. During this class we will be learning some of the laws of nature. In each case, we will examine this assumption; that life is shaped by and exists within the constraints of purely physical laws. We will ask whether living things ever evade the limits these laws place, as they might if life involved something beyond the physical; something metaphysical or supernatural. We will also ask how the extraordinary diversity we see in life could come about *as a result* of these physical laws.

There is one extremely important tool needed to understand the interplay between physical laws and life: evolution. The idea that life evolves through natural selection of random variations provides our only tool for understanding the diversity of life. Evolution has allowed life to find incredibly various and seemingly ingenious ways to function. Working within the limits provided by physical laws, evolution will allow us to understand why animals never acquired wheels, why cells are the size they are, why hummingbirds eat

¹ Oxford English Dictionary

several times their weight in food each day, and why the largest animals which have ever lived all swim in the sea.

What we'll accomplish in this course will only scratch the surface of this profound and important topic. But even a superficial look can teach you a lot about the inescapable unity between the physical and biological worlds. At a minimum you should learn how physical laws constrain organisms, and with luck this new understanding will change the way you think about life.

1.2 Three examples to set the stage

Let's start with a few examples to illustrate how life has evolved to work around the limits placed on it by physical laws. Let's start with one of the most obvious connections: size and shape.

The Spherical Cow and Modeling

There is a famous joke about cows and professions; told in many forms over the years. A dairy farmer is having trouble making ends meet, and hopes to find a way to enhance the productivity of her farm. For reasons lost in the mists of time, she calls in a psychologist to help.

The psychologist conducts a series of interviews with the farmer, her family, and the cows, and records their reactions to a variety of visual stimuli. In the end, he tells the farmer "Your cows are suffering from Ruthvenian post-lactic stress disorder. To enhance their productivity requires a more nurturing climate. Paint the walls of the barn a cool, neutral color, provide them with quietly energetic music, and be sure to give them a daily shiatsu massage." This doesn't work. Indeed the cows become very relaxed, but this only makes them rather harder to milk.

Next she turns to a biologist, who sends a graduate student to take cell cultures from inside each cow's mouth, has a team of lab technicians extract their DNA, and uses a room full of capillary electrophoresis machines to sequence it. His report to the farmer suggests supporting a new research program to splice the DNA of a Minke Whale into the cow, allowing them to grow much larger, produce more and richer milk, and not wander about the farm so much. The Minke Whale, after all, has the richest milk of all mammals. The farmer, imagining the bad press that would accompany this Franken-milk, politely thanks the biologist and sends him on his way.

Finally, the farmer calls on a physicist. Unlike the others, the physicist doesn't examine the cows or consult the farmer's family at all. Instead, she goes to the chalkboard, draws a large circle, says: "Assume each cow is a sphere" and begins to construct a quantitative, predictive model on the board...

The point of this joke, like so many, is to illustrate partial truths. Psychologists do look for answers in the minds of their subjects. Biology has found genetic research incredibly fruitful. And physicists achieve much of their success by building simple mathematical models. To motivate this seemingly strange approach, let's see where we can go with this "spherical cow approximation", which we will refer to from now on as the SCA.

If a cow were a sphere, life would be simpler. A spherical cow, unlike a real one, has a shape fully described by a single number; its radius. Tell me the radius of a spherical cow, and I can quickly calculate both its volume ($\frac{4}{3}\pi r^3$) and its surface area ($4\pi r^2$). For a real cow, with its complicated shape,

calculating volume and surface area is hard to do. What's the formula for the volume of a cow given its height at the shoulder h ? There isn't one. It's much easier to work with a spherical cow. But although these SCA answers are easy to calculate, they're also obviously **wrong**. I don't expect to precisely predict the mass of a cow from the SCA. What use is a model so simple that it ignores the shape of a cow? As it turns out, there are many things that such a model *can* accurately predict.

What would happen if I changed the size of a cow? How would its volume and surface area change? Without a model, we could only do the experiment: grow a big cow and measure it. But given the SCA, we can predict what will happen. For our sphere, volume is related to radius through the equation $V = \frac{4}{3}\pi r^3$. If we double the size of r , the volume increases by a factor of 2^3 , or 8. We can also predict how the surface area ($A_{\text{surface}} = 4\pi r^2$) changes: doubling the size will increase surface area by a factor of 2^2 , or 4.

Why might the farmer care? Milk production probably varies with volume. Doubling the size of the cow would yield 8 times as much milk. But what if she's really out to make leather? Doubling the size of the cow would yield 4 times as much hide. Now of course there are other implications of size. The amount of food and oxygen an organism needs is largely governed by mass. Each cell must be kept alive, and more mass means more cells. So doubling the size of a cow increases its food needs by a factor of eight! A farmer out to make leather would, as a result, generally prefer a lot of small cows to a few big ones. Each would have a relatively large surface area to volume ratio, and hence would produce leather relatively cheaply.

There is another way we might use this model. It allows us to predict how the surface area of an organism ought to vary with its volume. Deriving this is simple. First use the equation for volume to find the radius as a function of volume, then insert this into the surface area equation:

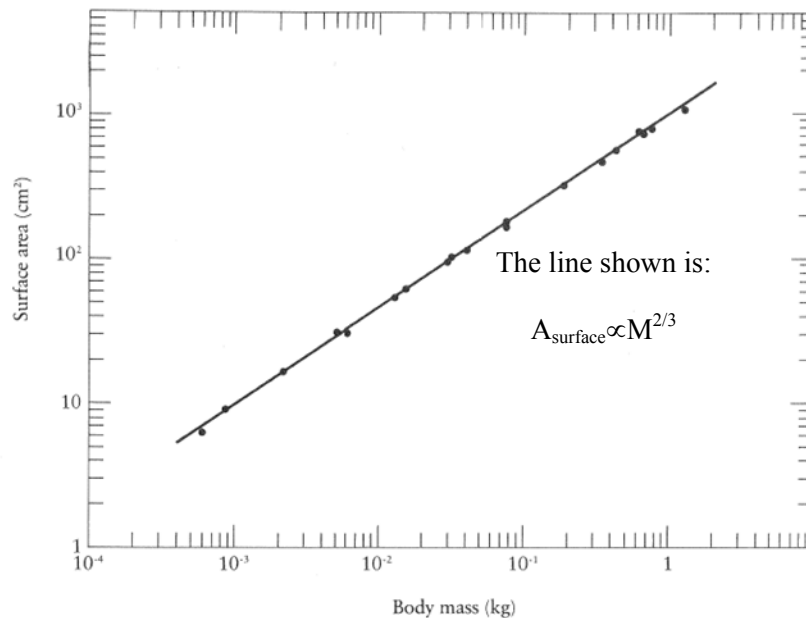
$$r = \sqrt[3]{\frac{3V}{4\pi}}$$

$$A_{\text{surface}} = 4\pi r^2 = (4\pi)^{\frac{1}{3}} 3^{\frac{2}{3}} V^{\frac{2}{3}}$$

This model leads us to suspect that surface area should be proportional to volume to the $\frac{2}{3}$ power. This relation, derived here for spheres, should apply to any organism that stays the same shape as it changes size. One good example is provided by a salamander called the dwarf siren (*Pseudobranchis striates*) which lives in the Southeastern United States. As it grows, it retains very much the same shape, providing a nice test of this model. Measurements of how its surface area varies with volume show, shown in the Figure below, illustrate just the relation we expect from this SCA¹.

We will make use of this and many other simple mathematical models of things, both living and not quite, often in this class. They will allow us to extract important facts from all the specific complexity of real biological circumstances. This kind of quantitative but clearly approximate modeling plays an important role in understanding life. We will return to it again and again in our efforts to understand the essentials of how physical laws enable and constrain life. We should note right from the start, however, that this approach can be perilous. Sometimes when we abstract out the details we lose the substance. Care,

caution, and an element of humility all play important roles in the generation of accurate and useful models.



Surface area vs. mass relation for the Dwarf Siren salamander

Convergent Evolution

Evolution, through random variation and subsequent selection of successive generations, allows life to find strikingly effective ways to work within the limits placed by physical laws. A beautiful example of this in action is provided by “convergent evolution”. When physical laws present a difficult problem for life, there are sometimes very few workable solutions. When this is so, the same solutions are often arrived at over and over, completely independently, during the course of evolution. While there are many examples of this which we will encounter in this course, perhaps the most visible, well documented, and delightful, is flight.

We all know what happens when we leave an object of any substance unsupported in the air: gravity pulls it to the ground. Sometimes the friction it feels when moving through the air slows its fall, but in the end what goes up must come down. To achieve the marvel of flight, to remain in the air for an extended period, climbing and diving at will, you must be able to generate forces large enough to overcome the gravity which pulls you down. As we will come to see later, when you want something to push up on you, you need only push down on it. Alone in the air, the only thing to push down on is the air itself. If you want to push hard on air, you need to push with something big; something like a wing. You can make flight easier by being as light as possible; reducing how hard gravity pulls you down. The best solution is clear: combine big wings with a body made as light as possible.

Life, through evolution, has found this solution at least four completely independent times, among the insects, reptiles, birds, and mammals. In each case the solutions are strikingly similar: large, thin, flexible wings attached to bodies with many adaptations designed to reduce weight.

Insects might seem an exception to the focus on reducing weight. After all, a big fat June Bug flies along just fine. How can this be? Recall the SCA. The volume of a creature increases like size^3 , while the surface area increases like size^2 . So long as shapes remain the same, the ratio of mass to wing area (volume to surface area) changes like $\text{size}^3 / \text{size}^2$, or like size. The bigger a flier is, the more important it will be for it to limit its weight and increase the relative size of its wings. The largest fliers, birds like condors and cranes, have quite enormous wingspans and surprisingly small masses. The California Condor has a wingspan of nearly three meters, but as mass of only about nine kilograms.

Insects live at the low end of this tradeoff, where the benefits of reducing weight are really not important. So there are lots of chubby insects, from beetles to bumblebees, which can still fly. Why are insects always pretty small? It turns out their sizes are not limited by the constraints of flight at all. We'll see what keeps them small in a moment. Throughout this course, we will see many other examples of convergent evolution; in the shapes of swimmers, the structures of eyes and ears, and in the ways in which organisms insulate themselves and prevent heat loss.

Diffusion and catching your breath

Most of the motion we associate with life seems willful: you throw a ball, a bird flaps its wing, a snail crawls across the floor. But one kind of motion, incredibly important for life, clearly doesn't require will: it just happens. This is transport on the molecular scale: what we would generally call "diffusion".

Imagine a rectangular box divided in two. The left hand side is filled with Nitrogen gas. In it huge numbers of N_2 molecules fly freely through space, colliding occasionally with one another or the walls. The right hand side is completely empty; a vacuum. Now suppose we remove the dividing wall. What will happen? Some of the molecules which would have hit the divider and bounced back will now just continue into the empty side, eventually reaching the far wall and bouncing back. After a while (a very short while indeed in this case) there will be essentially equal numbers of molecules on both sides of the box.

How did this happen? Did anyone decide to push those molecules across and *make them* spread out evenly? This kind of motion, which *just happens* as a result of random thermal motion, is called "diffusion". It happens because atoms and molecules are always moving with speeds that depend on their temperature. If you give things which rattle around randomly a chance to spread out, they will. Not because they want to, or because anything is pushing them; just by chance.

How does life use this? When it comes time to deliver a whole mess of stuff (like a mouthful of food or a big gulp of air) simple pushes and pulls will do the job. But if you need to deliver individual molecules to where they're needed, our familiar forces don't work; no tool you have can grab a single molecule and push it around. On this scale diffusion has to take over. Let's take a basic example: getting oxygen molecules into a cell. Imagine you put some air next to a cell. There will be oxygen in the air, and let's assume there is no oxygen inside the cell. If the cell membrane allows oxygen to pass through, then when

oxygen outside the cell hits the wall it will pass through. It will continue to build up inside the cell until there is, on average, just as much flowing out as flowing in.

How fast does this diffusion process happen? The rate depends on a number of things, including how much oxygen is on one side and the other, the temperature of the stuff on both sides, how much surface area there is to diffuse through, and how permeable the membrane is. A key factor here is surface area. Remember, the amount of oxygen something *needs* depends on its total volume, which changes like size^3 . The amount of oxygen *it can absorb* through diffusion depends on surface area, which changes like size^2 . So in our SCA, the amount of oxygen available for each little bit of an animal should change like surface area / volume, or like size^{-1} . This sounds like a losing game. The bigger you get, the less able you are to supply oxygen to your cells. As organisms get larger, it should become harder and harder for them to deliver oxygen to their cells.

How has evolution gotten around this limitation? The answer is to break the spherical cow approximation. A key part of the SCA is the assumption that when things change size, they stay the same shape. Our spherical cow is always a sphere. We call this kind of change in size “isomorphic”, meaning keeping the same shape. Evolution has overcome the challenges diffusion imposes by changing shape while changing size, and doing this in ways which make the surface area for oxygen exchange increase more rapidly than it would in the SCA.

For most “big” organisms, this is done by developing gills or lungs. Your own lungs, for example, are a spongy mass of more than a half-billion tiny sacks called alveoli. Their total surface area is around 100 square meters. Is this big? You’re about 2 m tall and 0.3 m wide, so your surface area (in skin) is perhaps 1.2 m^2 . By growing complex, spongy lungs, your body increases the available surface area for oxygen diffusion by a factor of 100, easily enough to let you grow big. Fish do the same trick by growing gills; huge numbers of very thin feathery sheets they expose to oxygenated water. If you wanted to deliver oxygen to your cells more directly, without the complication of squishy, disease-prone lungs, you’d have to be about 100x smaller. That would make you 2 cm tall; about the size of a cricket.

In the end, this is why insects have never become large. They mostly deliver oxygen directly to their cells through a system of narrow tracheal tubes. This system acts almost entirely through diffusion, and doesn’t allow them to really change shape while changing size. As a result, it provides one fundamental limit to the size of insects. Without something like lungs, they cannot grow much larger.

These three examples – modeling size and shape, thinking about convergent evolution, and seeing the constraints of diffusion – should give you a good flavor for how we will approach the application of physics to life in this course.

1.3 Some differences among the sciences and how you study them

Many of you will take courses in a variety of different sciences while preparing for a career. There are superficial aspects of practice in each discipline which combine to make physics, chemistry, and biology seem very different, obscuring the connections among them. There is one in particular you might want to watch out for from the very beginning.

Biology and chemistry make extensive use of unique, specially invented terminology, often derived from Latin or Greek. This choice of ancient languages preserved in part to support this penchant for specific

naming, makes clear the desire to avoid using “everyday” language for the naming of things. Specific names are adopted to emphasize as clearly as possible the differences among things, and to stress for example the distinction between deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). A significant part of the study of biology and chemistry is dedicated to learning all this terminology

Physics, by contrast, tends to focus on describing things in the simplest possible way. In so doing, it often picks up terminology from everyday life for use in its theoretical framework. Examples of physics terms include the normal force, friction, work, energy, force, action, pressure, tension, stress and strain, the big bang, the standard model, and string theory. Using this kind of language is nice because it helps to avoid the possible obscuring effects of unfamiliar terms. It can also be treacherous, because these familiar words bring with them everyday meanings different from those used in physics. To overcome this pitfall, you have to be particularly careful to notice when familiar words are adopted in physics with much more particular meanings. We will try to point this out, but you have to heighten your sensitivity to it as well.

There is another feature which tends to distinguish introductory physics courses from those in the other sciences. In this course, we will introduce you to a relatively small number of principles, and then ask you to apply them in a wide variety of situations. There is little to memorize, but much to practice. Rather than being asked to replicate things you have been told or shown, a physics course will ask you to apply principles you have learned in new ways. For example, you ought by now to be able to use the SCA to decide whether objects falling a long way through the air fall faster or slower when they are large.

1.4 Tools and skills for this course

To apply the laws of physics, there are a number of basic principles of mathematics which you will use regularly. The remainder of this first chapter provides a quick review of some of the most important. These are all tools we will use extensively, and rather than remind you of them every time they come up we’re going to put them all here.

Volume and Surface Area

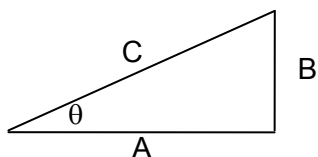
There are many simple shapes for which volume and surface area can be simply expressed:

Shape	Surface Area	Volume
Cube of edge length L	$6L^2$	L^3
Sphere of radius R	$4\pi R^2$	$4\pi R^3/3$
Cylinder with radius R and length L	$2\pi R^2 + 2\pi RL$	$\pi R^2 L$

For other shapes, like a wombat or an automobile, simple formulas for surface area and volume don’t exist. Remember the SCA though; there are scaling rules which always apply. If you change the size of an object, keeping the shape the same, the surface area will increase like the size² and the volume will increase like the size³. You can see this is true in the above trivial examples, and you can extend it to any shape you like by imagining the shape constructed of tiny cubes.

Trigonometry and the Pythagorean Theorem

As we discuss various geometric properties such as size, shape, motion, etc., you will have to use trigonometry in many basic ways. Here are the essentials which would be ready to apply. Given a right triangle with sides that have length A, B, and C, we can write the following:



$$A^2 + B^2 = C^2$$

$$\sin\theta = B/C = \text{opposite/hypotenuse}$$

$$\cos\theta = A/C = \text{adjacent/hypotenuse}$$

$$\tan\theta = B/A = \text{opposite/adjacent}$$

People often remember this with the mnemonic “SOHCAHTOA” (for sine = opp/hyp, etc.), and you may find that useful too. Here’s another useful thing to recall: angles can be measured either in radians (which run from 0 to 2π) or in degrees (which run from 0 to 360). You’ll need to be careful about which you’re using with your calculator, especially when inverting trigonometric functions to solve equations like

$$\sin(\theta) = 0.45$$

$$\theta = \sin^{-1}(0.45) = 26.74^\circ = 0.467 \text{ radians}$$

Some basic calculus

Calculus is a branch of mathematics dedicated to describing change. Physics is all about change; not about how things are, but about how they change. Calculus was invented, in large part by Newton and Leibnitz, as the central tool of physics. As a result, any serious understanding of physics requires reference to calculus. This course does not, however, require a very elaborate application of calculus. So while we will very often include the ideas of calculus in what we discuss, you won’t have to deploy the many methods of calculus very often.

You will need to understand that derivatives of functions describe their rates of change (their slopes), and that integrals of functions describe areas under them. We will do some simple calculus derivations occasionally, and you should be comfortable with understanding them.

Intensive and extensive quantities

In our study of physics we will often speak of the properties of objects: their masses, electric charges, forces applied to them, and so on. These quantities apply to particular objects. We will also often speak of quantities which are properties of materials, rather than the complete objects of which they are made.

A familiar example of such a so-called ‘intensive’ quantity is density. The average density of an object can be found by dividing its total mass by its total volume:

$$\rho_{average} = \frac{M_{total}}{V_{total}}$$

Other examples of intensive quantities include electric charge density, temperature, elasticity, and pressure. Intensive quantities will often vary in space, taking on different values at different locations. The density of an object, for example, is determined by what it's made of. Most living things (like you) are made of a mix of things (muscle, bone, brain). As a result, your density varies from place to place inside your body. Nevertheless, most animals are mostly made of water, so often you're not too far wrong if you use the density of water to estimate the density of an animal. Conveniently, water has a memorable density of about 1000 kg/m³.

How could you use this to estimate your mass? There's no formula to get your volume from your height. So let's estimate it by imagining you're a cylinder, say 1.8 m high and with a radius of 10 cm (about 4 inches). This would give us a volume of 0.056 m³ and a mass of about 56 kilograms. This is a nice example of how an intensive property of a material (the density) can be used to find an extensive property (the mass) of an object.

Estimation

This provides a nice introduction to the topic of estimation. We will often estimate things in this course. Why not just be precise, use equations, and calculate exact answers? There are at least two different reasons.

First, most of what happens in the world is incredibly complicated. This makes precise description, in the form of a perfect, tidy equation, impossible. Fortunately, this complexity doesn't leave us helpless to describe or unable to predict what will happen. It simply means we will have to approximate; to construct models which capture some of the most important features of the situation, while glossing over less significant details. Our spherical cow is a great example of this. It doesn't tell us the volume of a cow. But it does give us an idea of how that volume changes as a cow grows.

There is another important reason to estimate. Even if we had a perfectly precise, tidy analytic theory, we still may not perfectly know the parameters involved. For example, if we want to know the mass of a cow, we may not know its precise height or length. We may not know its density or detailed shape. What's the density of a cow? How could you estimate this? Well, like most land animals, cows can swim, a little at least. This means they can nearly float. This means their density must be close to the density of water, which is one of those nice, useful, numbers you should just know...

It is very useful, and important, to learn to make quantitative estimates in situations where perfect knowledge is absent. This appears in all arenas of life. If, for example, someone told you there were 5,000 piano tuners in Ann Arbor, should you believe them? Most of you have probably heard the five second rule: "if you drop a piece of food on the ground and pick it up in less than five seconds, it's OK to eat it". Is this nonsense or true? Why five seconds and not 2.5 or 10? What happens in five?

Units of physical quantities and their associated dimensions

Quantitative description of anything implies measurement: comparison to some standard of reference. Measurements discussed in this book will involve comparisons to just a few fundamental standards: time, distance, mass, and electric charge. In each case we will use the units defined as part of the ‘Système international d’unités’. These are usually just called SI units.

Time: seconds, currently defined as time required for a Cesium atom to vibrate 9,192,631,170 times

Distance: meters, currently defined as distance traveled by light in $1/299,752,458$ second, about 3.28 feet

Mass: kilograms, currently defined as mass of a little cylinder kept in Paris, about 2.2 pounds in more familiar units

Electric charge: coulombs, currently defined as the amount of electric charge contained by about 6.25×10^{18} electrons

Now anything you might measure, like a distance, or a time, might be measured in a variety of different units. Time, for example, might be measured in seconds, or hours, or days. Some particular period of time, 45 seconds say, actually has only one duration. We might *measure* it in many different units, but it’s always really the same thing. To convert this one period from one set of units to another, we can take advantage of conversion factors, multipliers which change units but have a numerical value of one:

$$45 \text{ seconds} * (1 \text{ minute} / 60 \text{ seconds}) = (45/60) \text{ minutes} = 0.75 \text{ minutes}$$

$$45 \text{ seconds} * (1 \text{ day} / 86400 \text{ seconds}) = (45/86400) \text{ days} = 5.2 \times 10^{-4} \text{ days}$$

Notice what we do in each case. Start with what you are given (45 seconds), then multiply by a “conversion factor”; a ratio of two times that are equal to one another, but measured in different units. Since the two are equal, the ratio is actually equal to one, and when you multiply by it, you leave the original time period unchanged. What the conversion factor does, then, is to change the *units* without changing the *value* of the measured quantity. Here are a few more examples:

$$1.8 \text{ meters} * (100 \text{ centimeters} / 1 \text{ meter}) = 180 \text{ centimeters}$$

$$56 \text{ kilogram} * (1 \text{ pound} / 0.454 \text{ kilogram}) = 123.4 \text{ pounds}$$

$$1 \text{ electron charge} * (1 \text{ Coulomb} / 6.25 \times 10^{18} \text{ electron charges}) = 1.6 \times 10^{-19} \text{ Coulombs}$$

Since we will work with a variety of different units, you will need to develop some facility with doing these conversions. Sometimes they will be more complicated. Let’s convert speed in meters per second to miles per hour:

$$1 \text{ meter/second} * (1 \text{ mile} / 1609 \text{ meter}) * (3600 \text{ second} / 1 \text{ hour}) = 2.24 \text{ mph}$$

With time we often define non-standard units imposed on us by circumstance. For living things on Earth the day governs one strong cycle of life, as does the year. These two essential units are unfortunately incommensurate: we now define each year to be 365.25 days. People in their social relations define other units, like the week. Our shorter time units, the second, minute, and hour, were originally defined in

reference to the day. Of course these units, though convenient for us, could only be appropriate on the Earth. They have no physically fundamental importance.

As we go through the class it may be useful to know how many seconds there are in a day or a year:

$$1 \text{ day} * (24 \text{ hours} / 1 \text{ day}) * (60 \text{ min} / 1 \text{ hour}) * (60 \text{ sec} / 1 \text{ min}) = 86,400 \text{ seconds}$$

$$1 \text{ year} * (365.25 \text{ days} / 1 \text{ year}) * (86,400 \text{ sec} / 1 \text{ day}) = 3.15 \times 10^7 \text{ seconds}$$

Notice that this latter is very close to $\pi \times 10^7$ seconds, which perhaps makes it a little easier to remember. Once we start counting in years, interesting coincidences emerge, like the fact that a typical 50 minute introductory physics lecture is just about 1 microcentury. Some lectures *seem* much larger, some much shorter. You can make of that what you will.

So that's what "units" are. What about "dimensions"?

When we ask what dimensions something has, we're asking about what kind of a quantity we're measuring, not how we're measuring it. Is it a distance, a time, a mass, a charge, or some combination of these? Notice that this is different from asking about units. When we talk about units, we're asking something more specific. How are we measuring this? What is it we're comparing this thing to? When we talk of dimensions, we're only asking about the nature of the thing we're measuring, not the particular system of comparison we use to measure it.

As it happens, there aren't so many fundamentally different dimensions of measurement in physics. For much of this course, we'll need just four:

Length: L

Time: T

Mass: M

Charge: Q

Other quantities of interest are measured in terms of these. For example, speed has dimensions of length divided by time (L/T), and density has dimensions of mass divided by volume (M/L³). Paying attention to the dimensions of things is important. The dimensions of a thing tell you what it really is. While the units can be changed by conversion factors, the dimensions cannot. There is no way to turn a length into a mass.

Orders of magnitude and scientific notation

In discussing the physics of living things we will need to talk about some things very far removed from everyday life. This will include objects ranging from the very small to the very large. In the spatial dimension we will talk about things ranging in size from atoms (with a typical radius of 0.000,000,000,1 m) to the Earth (with a radius of 6,400,000,000 m). In time we will consider the time it takes two atoms to bond (only about 0.000,000,000,001 s) to the age of the universe (more like 4,400,000,000,000,000 s). In

mass, we will ponder both electrons (very light at 0.000,000,000,000,000,000,000,000,000,091 kg) and, again, the Earth (a rather more massive 6,000,000,000,000,000,000,000 kg).

Dealing with a large range of scales is mentally challenging; grasping the large and small requires enormous imagination. Fortunately, science has invented a tool, scientific notation, which provides helpful crutch. It can be used to precisely discuss an enormous range of parameters in a tidy, concise way. We will often use scientific notation and the Greek prefixes associated with it. The primary ones we will use are:

Tera-	10^{12}	a trillion
Giga-	10^9	a billion
Mega-	10^6	a million
Kilo-	10^3	a thousand
Centi-	10^{-2}	1/100
Milli-	10^{-3}	1/1000
Micro-	10^{-6}	1/1 million
Nano-	10^{-9}	1/1 billion
Pico-	10^{-12}	1/1 trillion

So when we speak of a centimeter, we mean 1/100 of a meter, and when we talk about a kilogram, we mean 1000 grams. You will have to always exercise care with these prefixes. They are the source of many student errors. You don't, for example, want to confuse micro and mega...

As our study of the universe has advanced, science has pushed these limits, and so there are now more official prefixes than there used to be. Going up, the next few are Peta-, Exa-, Zeta-, and Yotta- (10^{15} , 10^{18} , 10^{21} , and 10^{24}), going down they are Femto-, Atto-, Zepto-, and Yocto- (10^{-15} , 10^{-18} , 10^{-21} , and 10^{-24}). These remain a bit specialized, but you should certainly know all the prefixes from Pico- to Tera- by heart.

Notice that just about all of these are multiples of a thousand. The exception (Centi-) is useful because the basic distance, a meter, is a quite a bit bigger than our hands, and many things we work with are the sizes of our hands or smaller. The next scale down (the millimeter) is a little too small for everyday things. So the occasional use of Centi- is an accident of convenience, a consequence of the size and focus of interest of humans.

You will need to learn to manipulate these things, so if you are rusty with exponents, I suggest you review them. Here are the basics:

$$10^a \times 10^b = 10^{a+b} \quad \text{so } (5 \times 10^8) \times (4 \times 10^9) = 20 \times 10^{17} = 2 \times 10^{18}$$
$$10^a / 10^b = 10^{a-b} \quad \text{so } (5 \times 10^8) / (4 \times 10^9) = 5/4 \times 10^{-1} = 1.25 \times 10^{-1} = 0.125$$

Scientific notation is a tool which allows us to talk about an enormous range of sizes, times, and masses. There are about as many nanoseconds in a single second as there are seconds in your life; each is an almost unimaginably tiny period. And yet physicists studying life now routinely produce pulses of laser light which are "femtoseconds" long. There are a million femtoseconds in each nanosecond, and hence 10^{15} , a thousand trillion, in every second. Scientific notation makes working with such enormous numbers tractable, but it isn't much help in our effort to more deeply understand them.

Each time you encounter the very large and very small, you should expend some imaginative effort on it. Your body contains approximately 10^{13} cells. Each contains a complete copy of your genome, with all 2.85×10^9 base pairs. Among your cells, there are about 10^{11} neurons, connected to one another through about 10^{14} synapses. The biggest neurons have masses of about 10^{-9} kilograms. Getting beyond your own skin, there are nearly 7×10^9 people on the Earth. Putting these together, there are now about a mole of neural synapses working on the planet. Wrapping your mind around numbers like these is probably a life-long endeavor, and is surely one of the pleasures of learning science.

1.5 The nature of things we might measure: scalars and vectors

When we set about quantifying the world, measuring things about it, we discover that not all things can be described in quite the same way. Many things we might like to measure are rather simple; they can be represented by a single number. A baseball has a mass. Everything there is to know about the baseball's mass is represented by a single number: 5.25 "ounces avoirdupois", or about 149 gm (according to the official rulebook). It also has a circumference (officially "not less than nine nor more than 9 1/4 inches", or about 23 cm). Just one number tells you everything there is to know about its circumference.

Physical properties which can be represented by just a single number are known as **scalars**. They are quite common. In addition to mass and diameter, they might include temperature, density, pressure, metabolic rate, pH, age, or even cost. All scalars properties can be fully described by just one number. We sometimes say that scalars are properties of things which have only a **magnitude**.

Some things we want to measure, especially in physics, are more complex. If we want to describe the wind for a sailor, it is not enough to simply provide its speed. To usefully describe the wind for a sailor, we need also to provide its direction. Another quantity like this is a force. To fully describe a force, to tell you everything you need to know about it, we have to give both its magnitude and its direction. There are many examples in physics, including force, displacement, velocity, and acceleration. Later in the course we will encounter electric, magnetic, and gravitational fields. All of these are called **vector** quantities. Vectors are things which require us to specify both a **magnitude and a direction** to give a complete description.

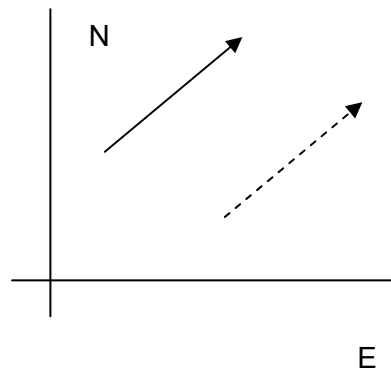
The point is to draw your attention to an essential difference. If you're analyzing something and the answer you seek is a scalar, you need only determine its magnitude. But if the answer to your question is a vector, you will have to determine both its magnitude and its direction. Without providing both, you cannot fully describe a vector.

Displacement as an example vector:

Scalars are pretty familiar things, so they don't need much further introduction. Vectors are considerably less so, as they were invented for and are largely used in physics. So we will take some time to talk about what vectors are and how we add, subtract, and multiply them.

Let's take as our example a displacement; a kind of instruction for a trip. To describe a trip, we have to say how far to travel, and also what direction to go. One way to do this is to specify the magnitude of the vector, and describe its direction by measuring an angle relative to some reference direction. Here's an example; you receive an instruction telling you to travel 30m in the direction North-East.

We could represent that trip graphically as a little arrow that looks like this.



What does this graph represent? Each of the axes represents position. The horizontal axis measures how far East or West you are from the (arbitrarily selected) origin, while the vertical axis measures how far North or South you are from this same arbitrary origin. The solid arrow shows a displacement “30m in the NE direction”. The dashed arrow also shows a displacement “30m in the NE direction”. Since these two displacement vectors have exactly the same magnitude and direction, they are precisely equal to one another. This is an important point. Vectors are not tied to particular points in a space. They don't say go from this particular spot to that; they just tell you how far to go and in what direction.

The reason for this is actually rather deep. The way things move can't be affected by how we choose to draw our coordinate system. If they were we could never know what was going to happen until after we defined a coordinate system. Displacement vectors like these will prove very useful in describing physics, and to be physically meaningful they have to be independent of particular starting and ending points.

There are a few small subtleties to consider. Usually, we will describe a displacement as having a positive direction: for example 10 meters North. Occasionally, it is convenient to speak of vectors with *negative* magnitude. Doing this implies that the vector has a direction opposite the one stated. So if we spoke of a vector which has a magnitude of -10 meters in the North direction, it would be identical to a vector with magnitude +10 meters in the South direction.

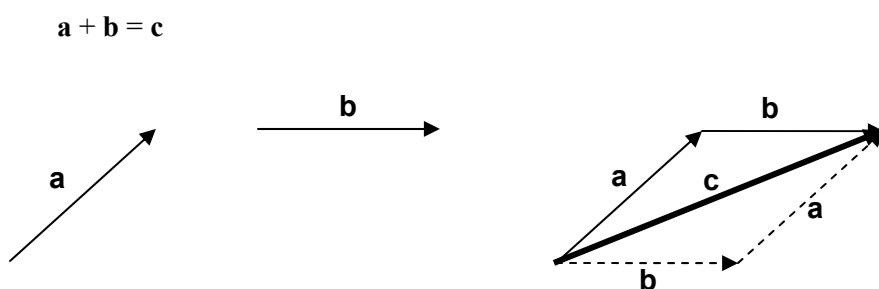
Displacement is the archetype of a vector. In this text, we will denote vectors by using boldface, so that while s might be a distance, \mathbf{s} is a vector. Another way to denote vector quantities is to draw the

appropriate symbols with little arrows over them, so we might use the symbol ‘ \vec{s} ’ is a displacement vector.

Adding and subtracting vectors

There are several ways we might want to manipulate vectors. The first is addition. Why might we want to add vectors? If we take two trips, we undergo two displacements, and the sum of the two is equivalent to taking some other single trip. Likewise, the sum of two forces is the same as a single equivalent force. So adding vectors really amounts to finding a single vector which is the equivalent of the combination of several other vectors.

Rather like scalars, the sum of two vectors **a** and **b** is equal to a single third vector, **c**, which is equivalent to doing **a** then doing **b**. So we can plausibly write:



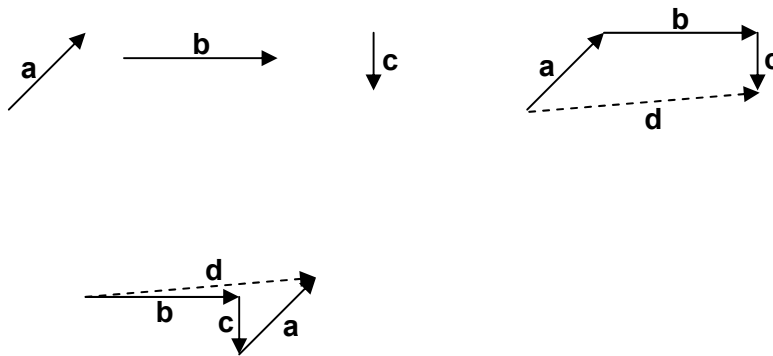
The order doesn't matter, which if you remember your fifth grade math, implies that vector addition has a property called commutativity:

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

In the figure we see the first way to discover the sum of two vectors. This is called graphical addition, or the ‘tip to tail’ method. It relies on the fact that vectors are NOT tied to particular points; they only have magnitudes and directions. Because of this, you’re free to move them around and line up the tip of the first with the tail of the second.

Note that we could go a step further, and consider what happens if we add three vectors together:

$$\mathbf{a} + \mathbf{b} + \mathbf{c} = \mathbf{d}$$



Notice from this example that the **same** resultant vector **d** is produced whether I take:

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} \quad \text{or} \quad (\mathbf{b} + \mathbf{c}) + \mathbf{a}$$

This different kind of independence from order is called associativity.

Multiplication of a vector by a scalar:

A vector is a quantity specified by both a magnitude and a direction. A "scalar" is something specified by just a magnitude. So a distance "3m" is a scalar, and a displacement like "3m Northwest" is a vector. It is possible to multiply a vector by a scalar. To do this we just multiply the magnitude of the vector times the scalar number, leaving the direction unchanged.



Vector Subtraction:

Imagine I have executed a displacement. If I want to execute a second displacement which will eliminate the effect of the first, what new displacement must I execute?

Just as

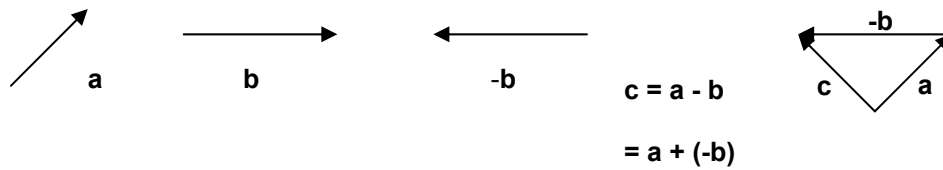
$$5 + (-5) = 0$$

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$$

where **(-a)** refers to a vector with the same magnitude as **a**, but the opposite direction, and **0** refers to a "null vector" of magnitude zero. This suggests how we should do vector subtraction:

$$\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})$$

which can be drawn graphically as:

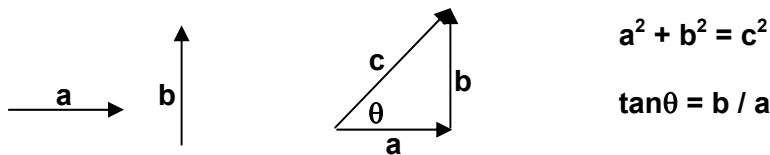


This emphasizes the fact that two vectors are equal if their magnitudes are equal, **and** their directions equal. When this is so, you can add the negative of one to the other and obtain the null vector:

$$\mathbf{a} + (-\mathbf{b}) = \mathbf{0}$$

Components of vectors, orthogonal motions:

A particularly illustrative example is the addition of two vectors which are perpendicular to one another. In this case, simple rules of trigonometry can be used to find the magnitude and direction of the sum.

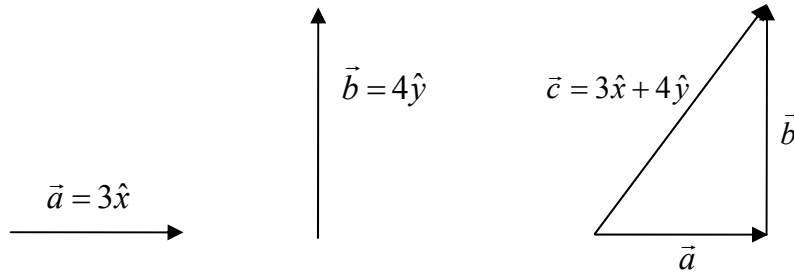


This example suggests what turns out to be an extremely useful way of thinking about vectors. A vector **c** can always be thought of as "made up of" the sum of two appropriate vectors **a** and **b**. We need only require $\mathbf{a} + \mathbf{b} = \mathbf{c}$. In this case, **c** is in every way equal to this sum. The two things, **c**, or $\mathbf{a} + \mathbf{b}$ are exactly the same. If we do this while requiring that **a** and **b** are perpendicular to one another this is called "resolving **c** into components". We will do this very often with vectors; it makes many vector calculations much simpler.

Very often, a notational simplification is also made. If I set up a simple x-y coordinate system, I can define a "unit vector" for each direction. Each of these unit vectors points directly along the axis it corresponds to and has a length of one in the units of choice; hence the name unit vector. Usually unit vectors are written using the name of the axis, either as a bold vector or with a little "hat" symbol:

$$x \text{ unit vector} = \mathbf{x} \text{ (or } \hat{x}\text{)} \quad y \text{ unit vector} = \mathbf{y} \text{ (or } \hat{y}\text{)}$$

To talk about the unit vector in the x direction we would say "x-hat". I can use these unit vectors to rewrite the vectors above:



Initially, breaking vectors into components doesn't seem like much of a help; why would we want to replace one vector with two? Wouldn't that just complicate things? There are several reasons why this can make life easier.

1. The first is mostly practical; it is often easier to work with vectors which are broken into components than it is with the original vectors.
2. The second is somewhat deeper. It is often the case that the motion of an object along one direction is completely independent of the motion along another; the physics of the problem can "decouple" these two "orthogonal" motions. When this is the case, it is often advantageous to break vectors into components because this emphasizes the important features of the motion, hence making it easier to understand.

Let's see how component notation can simplify vector manipulation. Consider the following example of the sum of three vectors. We'll use the unit vector notation **E** for East and **N** for North.

Add three vectors:

$$\mathbf{a} = 5\text{m East} = 5\mathbf{E}$$

$$\mathbf{b} = 8\text{m North} = 8\mathbf{N}$$

$$\mathbf{c} = 6\text{m } 30^\circ \text{ East of North} = (6 \cdot \sin 30)\mathbf{E} + (6 \cdot \cos 30)\mathbf{N} = 3\mathbf{E} + 5.2\mathbf{N}$$

so the sum is:

$$\mathbf{a} + \mathbf{b} + \mathbf{c} = (5 + 3)\mathbf{E} + (8 + 5.2)\mathbf{N} = 8\mathbf{E} + 13.2\mathbf{N}$$

We can work out the magnitude and direction of this final vector in the way we did for adding perpendicular vectors above:

$$\text{Magnitude } m^2 = 8^2 + 13.2^2 \quad \text{or} \quad m = 15.4\text{m}$$

$$\text{Direction } \tan \theta = \text{opp/adj} = 13.2 / 8 = 1.65$$

$$\text{Or} \quad \theta = \arctan(1.65) = 1.02 \text{ radians} = 58.8^\circ$$

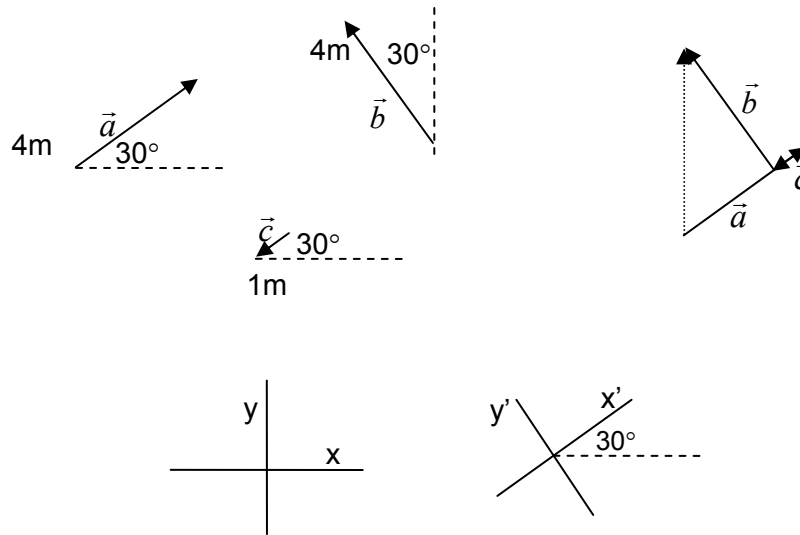
Notice that this would have been just as easy if there were 30 vectors instead of three. So when you have to add or subtract vectors, it is usually easiest to do it by components.

Picking the right coordinate system:

Now often you can *greatly* simplify a problem by using some feature of the arrangement of elements in a problem to simplify its solution. When we talk about "picking the right coordinate system" for a problem, this is usually what we mean. A couple of examples will give the general idea:

First a simple one: Add two displacement vectors 4m NE and 3m SW. We could break this into N and E components, but it is easier to add them along the direction NE/SW. Then we immediately find that their sum is a vector 1m NE.

Here is a second, slightly more complicated example:



We could resolve this into components along horizontal and vertical x and y axes, but that would be hard. Finding the sum will be easier if we think about a coordinate system rotated 30° counterclockwise. This is shown as the x' - y' coordinate system in the figure.

Writing vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} in components along these x' - y' axes is simple:

$$\mathbf{a} = 4\mathbf{x}'$$

$$\mathbf{b} = -1\mathbf{x}'$$

$$\mathbf{c} = 4\mathbf{y}'$$

So the sum of the three in this coordinate system is just:

$$\mathbf{r} = 3\mathbf{x}' + 4\mathbf{y}'$$

Which is a vector with magnitude $|\vec{r}| = \sqrt{3^2 + 4^2} = 5$, and direction $\theta = \tan^{-1}(\text{opp/adj}) = \tan^{-1}1.33 = 53.1^\circ$.

Does it matter physically that we've defined this vector in an unusual coordinate system? Not at all. It's always the same vector no matter how we choose to measure it.

Components and vector equality

We say that two vectors are equal when both their magnitudes *and* their directions are the same. It's also true that if any two vectors are equal, each of their individual components (along the x, y, and z axes for example) must be equal. So if we have two vectors **A** and **B** and they're equal, we can write:

$$\mathbf{A} = \mathbf{B} \quad \text{or} \quad A_x = B_x \quad \text{and} \quad A_y = B_y \quad \text{and} \quad A_z = B_z$$

This alternate way of writing things will often be simpler to keep track of than the more general definition of vector equality. So a lot of times when we know two vectors are equal we'll go ahead and write out three independent equations, one for each component. Since the equations for each component are just scalar equations, they are usually much simpler to work with.

Velocity vectors:

So now we have displacement vectors and we have some ideas about how to manipulate them. Apparently velocities must also be described with vectors; to completely specify them we need to know both how fast things are going and in what direction. We can define a velocity vector, averaged over some period of time Δt , in a straightforward way from the displacement vector:

$$\vec{v}_{avg} = \frac{\Delta \vec{s}}{\Delta t}$$

As we shrink the length of the period of time Δt over which we average, we determine this velocity over an infinitely short period of time, and speak instead of the instantaneous velocity:

$$\vec{v}_{instantaneous} = \lim_{\Delta t \rightarrow 0} \left(\frac{\Delta \vec{s}}{\Delta t} \right) = \frac{d\vec{s}}{dt}$$

Notice carefully what this is. The velocity vector is really just a scaled version of the displacement vector. In other words it is just the displacement vector multiplied by a scalar number; the inverse of the time it took to make this displacement ($1/\Delta t$). What this means is that the velocity vector **always** points in the same direction as the displacement vector.

Because motion takes place in three spatial dimensions, many things we will use to discuss motion this semester will be vectors; including forces, accelerations, stresses, flow rates, etc. It is important that you understand vectors very clearly, and that's why we're expending effort on them now.

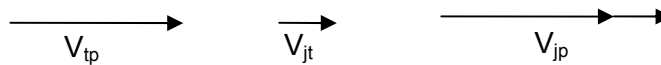
1.6 Decoupled motions and vector components:

OK, so looking at vectors by their components is a useful convenience, a nice simplification of some problems. It is also useful because the motion of objects along different directions can often be

independent, so that if we break it into components we can consider each motion independent of the others. One particularly nice example of this is the idea of relative velocity.

The "relative" we're talking about here is the constant velocity motion of some object observed by two different observers, who are themselves moving relative to one another. Start with a simple example:

Joe is on a train moving past the platform at 4m/s. He walks forward in the train with a speed of 1m/s *relative to the train*. What is Joe's speed relative to the platform? In vector form this problem can be written in a simple way:



V_{tp} = speed of train relative to platform = 4m/s

V_{jt} = speed of Joe relative to the train = 1 m/s

V_{jp} = speed of Joe relative to the platform = $V_{jt} + V_{tp} = 4 \text{ m/s} + 1 \text{ m/s} = 5 \text{ m/s}$

It's fairly obvious how this works when both motions are along one direction. How can we use the same approach when they are not? The following example gives the idea.

A boat can travel at 3 m/s through the water. It steers straight across a river which flows past the



shore at 5 m/s. What is the velocity of the boat relative to the shore?

So the magnitude of the boat's velocity relative to the shore (\vec{v}_{BS}) is:

$$|\vec{v}_{BS}| = \sqrt{(3 \text{ m/s})^2 + (5 \text{ m/s})^2} = 5.8 \text{ m/s}$$

And its direction is $\theta = \tan^{-1}\left(\frac{5 \text{ m/s}}{3 \text{ m/s}}\right) = 59^\circ \text{ W of N}$.

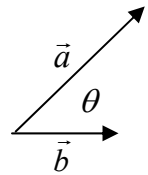
1.7 Vector Multiplication:

We have seen above how to multiply vectors by scalars already; you simply multiply the magnitude of the vector by the magnitude of the scalar. How do we multiply vectors with vectors? There's no *a priori* obvious way we should define vector multiplication, but as it turns out there are two physically *useful* ways to do it. The first produces a scalar as the product of two vectors, and the second produces a vector as the product of two vectors. We consider each in turn below.

The scalar product: $\mathbf{a} \cdot \mathbf{b}$

The 'scalar product' of two vectors produces a scalar, just a number. It is defined so that the number produced expresses the degree to which the two input vectors are aligned with one another. The formal definition is:

$$\vec{a} \cdot \vec{b} = |\vec{a}| |\vec{b}| \cos(\theta)$$



where θ is the angle between the two vectors. Since $|\vec{a}| \cos(\theta)$ is the component of \vec{a} along \vec{b} , and $|\vec{b}| \cos(\theta)$ is the component of \vec{b} along \vec{a} , there are two ways interpreting the scalar product defined in this way. It is either:

The component of \vec{a} along \vec{b} , times the magnitude of \vec{b}

or

The component of \vec{b} along \vec{a} , times the magnitude of \vec{a}

In either case, the scalar product is a kind of "colinear product" or a product of the colinear parts of a pair of vectors. Because the little 'dot' symbol is used to denote this operation, it is sometimes called 'the dot product'.

What kinds of questions will we use the scalar product for? Eventually we will want to keep track of how much objects move up and down. Imagine a bird flies through a three dimensional displacement vector that we can write as a vector \vec{d}_{bird} . If we want to know how much higher the bird is at the end of this displacement compared to the beginning we could use the scalar product. If we 'take the dot product' of this displacement with a unit vector which points straight up. Let's call this unit vector \hat{y} . If we do this, we can write:

$$\text{Distance the bird rises} = \Delta y = \vec{d}_{bird} \cdot \hat{y}$$

Likewise, this gives a shorthand for finding the components of a vector \mathbf{v} :

$$v_x = \vec{v} \cdot \hat{x}$$

$$v_y = \vec{v} \cdot \hat{y}$$

$$v_z = \vec{v} \cdot \hat{z}$$

where \hat{x} , \hat{y} , and \hat{z} are unit vectors in the x, y, and z direction.

The vector product: $\mathbf{a} \times \mathbf{b}$

The vector product takes two vectors and makes a third, new vector out of them.

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

where the magnitude of \mathbf{c} is given by:

$$|\vec{c}| = |\vec{a}| |\vec{b}| \sin(\theta)$$

and the direction of \vec{c} is perpendicular to the plane defined by \vec{a} and \vec{b} in a direction given by the right hand rule. The right hand rule says you should:

- Take your right hand
- Point your fingers in the direction of the first vector (\vec{a} in this case)
- Turn your hand until you can "curl" your fingers in the direction of the second vector (\vec{b})
- Now your thumb defines the direction of the vector \vec{c} .

From this definition you can see that the vector \mathbf{c} is always perpendicular to both \vec{a} and \vec{b} . The vector product is a kind of measure of the amount of perpendicularity of two vectors. Because the symbol 'x' is used to denote this operation, it is often called 'the cross product'.

Note that this vector product has the special property that it does not commute. That is:

$$\vec{a} \times \vec{b} \neq \vec{b} \times \vec{a}$$

in fact it "anticommutes"

$$(\vec{a} \times \vec{b}) = -(\vec{b} \times \vec{a})$$

Where will we use the vector product in physics? One good example has to do with rotation. If you want to get something to start rotating, you must apply a force to it. The ability of the force you apply to make the object rotate depends on both where you apply the force and in what direction you push. First we define a radius vector \mathbf{r} which goes from the center of rotation (the hinge of a door for example) to the point where the force is applied. Given this vector \vec{r} and the force vector \vec{F} , we will quantify the 'ability of this force to create rotation' by defining the torque $\vec{\tau}$ with the vector product:

$$\vec{\tau} = \vec{r} \times \vec{F}$$

Don't worry if this is confusing now. It's just an example which you ought to recognize when we return to it later.

Multiplying vectors using components

It's often the case that you'll have two vectors written in terms of components. For example, you might have:

$$\vec{a} = a_x \hat{x} + a_y \hat{y} + a_z \hat{z}$$

$$\vec{b} = b_x \hat{x} + b_y \hat{y} + b_z \hat{z}$$

Once you have expressed the two vectors in this way, you can multiply the vectors in either the scalar or the vector product in a convenient way. This approach takes advantage of the fact that these operations are distributive, and that the scalar and vector products of the unit vectors are simple.

For the scalar product we have:

$$\hat{x} \cdot \hat{x} = \hat{y} \cdot \hat{y} = \hat{z} \cdot \hat{z} = 1$$

while:

$$\hat{x} \cdot \hat{y} = \hat{x} \cdot \hat{z} = \hat{y} \cdot \hat{x} = \hat{y} \cdot \hat{z} = \hat{z} \cdot \hat{x} = \hat{z} \cdot \hat{y} = 0$$

You should be able to see why this must be true. Remember the scalar product measures colinearity. Two identical unit vectors are perfectly collinear. Two perpendicular unit vectors are not collinear at all.

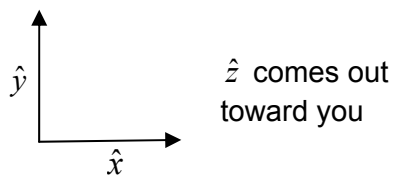
What about the vector products of unit vectors? Here the opposite is true. The vector product measures something about how perpendicular vectors are. The vector product of two identical unit vectors is zero; they aren't perpendicular at all. The vector product of two perpendicular unit vectors has magnitude of one, but now it's a new vector. In fact it's a unit vector in the third direction! In particular:

$$\hat{x} \times \hat{x} = 0 \quad \hat{x} \times \hat{y} = \hat{z} \quad \hat{x} \times \hat{z} = -\hat{y}$$

$$\hat{y} \times \hat{x} = -\hat{z} \quad \hat{y} \times \hat{y} = 0 \quad \hat{y} \times \hat{z} = \hat{x}$$

$$\hat{z} \times \hat{x} = \hat{y} \quad \hat{z} \times \hat{y} = -\hat{x} \quad \hat{z} \times \hat{z} = 0$$

If you draw a little coordinate system you should be able to use the right hand rule to check all the elements of this little table.



We can illustrate the utility of vector multiplication with components by providing a few examples. Let's look at the scalar product first:

$$\vec{a} \cdot \vec{b} = (a_x \hat{x} + a_y \hat{y} + a_z \hat{z}) \cdot (b_x \hat{x} + b_y \hat{y} + b_z \hat{z})$$

$$\vec{a} \cdot \vec{b} = a_x b_x \hat{x} \cdot \hat{x} + a_x b_y \hat{x} \cdot \hat{y} + a_x b_z \hat{x} \cdot \hat{z} + a_y b_x \hat{y} \cdot \hat{x} + a_y b_y \hat{y} \cdot \hat{y} + a_y b_z \hat{y} \cdot \hat{z} + a_z b_x \hat{z} \cdot \hat{x} + a_z b_y \hat{z} \cdot \hat{y} + a_z b_z \hat{z} \cdot \hat{z}$$

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

which is pretty simple. So here's an example:

$$\vec{a} = 3\hat{x} - 8\hat{y} + 5\hat{z} \quad \vec{b} = 2\hat{x} + 1\hat{y} + 3\hat{z}$$

What's the dot product? This would be hard to guess or execute graphically, as they're both in 3D. In component form it's simple:

$$\vec{a} \cdot \vec{b} = (3 \times 2) + (-8 \times 1) + (5 \times 3) = 13$$

That's it! Notice that the answer here is a scalar, as it should be for the scalar product.

The component approach to vector multiplication works the same way for the vector product. Here is a basic example for this:

$$\vec{a} = 3\hat{x} - 8\hat{y} \quad \vec{b} = 1\hat{y} + 3\hat{z}$$

What's the vector product of these two? Just expand it out and use the table above to fill in the appropriate cross-products of unit vectors.

$$\vec{a} \times \vec{b} = 3(\hat{x} \times \hat{y}) + 9(\hat{x} \times \hat{z}) - 8(\hat{y} \times \hat{y}) - 24(\hat{y} \times \hat{z})$$

$$\vec{a} \times \vec{b} = 3\hat{z} - 9\hat{y} - 24\hat{x}$$

You should notice that in this case the answer is itself a vector.

1.8. Life's media: air and water

Throughout this book we will consider how the laws of physics enable life and looking for ways in which they limit what life can do. Life on Earth is incredibly diverse, taking a still unenumerated variety of forms and making its way everywhere from the top of the atmosphere to the bottom of the ocean, and even some distance within the Earth. While the conditions of life vary from hotter than boiling to much colder than freezing, all of life exists within, and is largely made of, two fluids: water and air. As we

explore the physical aspects of life, we will continually find the properties of these two fluids playing a central role. So we take a moment here to begin introducing these essential media of life.ⁱⁱ

Both air and water are fluids; they can flow, rearranging their shapes to flow through constrictions and around objects. This malleable nature is essential. It allows the transport of nutrients into and wastes out of living things, bringing new things to and from the boundaries of organisms. It also allows us to move relative to our surroundings. As we will see, the motion of air and water around living things and their motion through these fluids are intimately related.

The air and water living things encounter are also chemically complex. While each has a predominant component (N_2 and H_2O), both mix freely with many other atoms and molecules. This chemical richness provides life with opportunities, all of which are put to good use by some organism or another.

Physically, the most obvious differences between air and water come from the fact that air is a gas while water is a liquid. Gasses are made of atoms and molecules which spend most of their time flying freely through space, unattached to one another and not interacting at all. They spend only a tiny fraction of their time colliding, though these collisions allow the gas to share energy efficiently, apply pressure to objects it encounters, and expand to fill whatever contains it. As a result, the density of a gas can vary enormously, and depends on how it is contained.

We can get a sense of what sets the density of air by using the ideal gas law as a kind of SCA model for its behavior. This law relates the temperature T , pressure P , and volume V occupied by a specified quantity of gas, in this case a number of moles n :

$$PV \propto nT$$

The product of pressure times volume is proportional to the amount of gas times the temperature. Rearranging this, we can make a prediction for how density ρ should change with pressure and temperature:

$$\rho \propto \frac{n}{V} \propto \frac{P}{T}$$

What does this mean for the air life lives in? The pressure of the air in the atmosphere is generated by the weight of all the air above a certain point. At the Earth's surface, it remains about the same, varying by about 10% as the weather changes. When the pressure rises, the air density increases, when it falls, it decreases. The temperature of the air is determined by the exchange of heat with its surroundings, and varies across the Earth's surface more dramatically than pressure, changing by as much as 50% from the Antarctic to the Sahara. When the temperature increases, the density of the air decreases.

While the density of the air encountered by life varies quite a bit, for standard conditions it has a rather small value of about 1.2 kg/m^3 . This low density, and the minimal way in which atoms and molecules in a gas interact, makes moving through air quite easy.

Liquids like water are much less free. In a liquid, atoms and molecules cling to one another jealously, never letting any one get far from the others. Since the atoms remain in intimate contact, never spreading out more or moving much closer together, the density of a liquid is much better defined. For sea water,

the most common kind on Earth, the typical value is about 1025 kg/m^3 . For fresh water the density is less, about 1000 kg/m^3 .

The density of water also varies with temperature, pressure, and composition. Fresh water is a few percent less dense than salt water. At most temperatures, water expands when heated and contracts when cooled, though again, only by a few percent. The density of water increases as you increase its pressure as well. But since the molecules in water are already pretty completely packed together, it is very difficult to change water's density in this way. Going from sea level to the deepest part of the ocean, the pressure on the water increases by a factor of more than a thousand, but the density of the water still increases by only a few percent.

Moving through water is quite a bit more difficult than moving through air. Not only does it have much more mass to move out of the way, the constant interaction of its molecules gives it a viscosity, a kind of sticky resistance to flow, which is much greater than that of air.

So here are the first facts to remember about air and water. The density of water is much greater than the density of air, about a thousand times greater. The density of air is about 1.2 kg/m^3 , of water more like 1000 kg/m^3 .

A Quick Summary of Some Important Relations

Scaling relations:

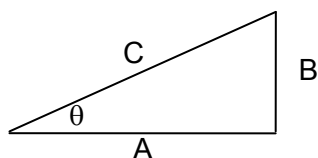
$$V_{\text{sphere}} = \frac{4}{3} \pi r^3$$

$$SA_{\text{sphere}} = 4\pi r^2$$

$$V_{\text{anything}} \propto (\text{size})^3$$

$$SA_{\text{anything}} \propto (\text{size})^2$$

Trigonometry:



$$A^2 + B^2 = C^2$$

$$\sin\theta = B/C = \text{opposite/hypotenuse}$$

$$\cos\theta = A/C = \text{adjacent/hypotenuse}$$

$$\tan\theta = B/A = \text{opposite/adjacent}$$

Units, dimensions, and magnitudes:

Mass	Kilograms	M
Length	Meters	L
Time	Seconds	T
Electric Charge	Coulombs	Q

Tera-	10^{12}
Giga-	10^9
Mega-	10^6
Kilo-	10^3
Milli-	10^{-3}
Micro-	10^{-6}
Nano-	10^{-9}
Pico-	10^{-12}

Vector operations:

The best general way to work with vectors is in component notation.

$$\vec{a} = a_x \hat{x} + a_y \hat{y} + a_z \hat{z}$$

$$\vec{b} = b_x \hat{x} + b_y \hat{y} + b_z \hat{z}$$

$$\vec{a} + \vec{b} = (a_x + b_x) \hat{x} + (a_y + b_y) \hat{y} + (a_z + b_z) \hat{z}$$

$$\vec{a} - \vec{b} = (a_x - b_x) \hat{x} + (a_y - b_y) \hat{y} + (a_z - b_z) \hat{z}$$

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z = |\vec{a}| |\vec{b}| \cos(\theta)$$

$$\vec{a} \times \vec{b} = (a_y b_z - a_z b_y) \hat{x} + (a_x b_z - a_z b_x) \hat{y} + (a_x b_y - a_y b_x) \hat{z}$$

$$|\vec{a} \times \vec{b}| = |\vec{a}| |\vec{b}| \sin(\theta)$$

Remember that the dot product produces a scalar (just a number), while the cross product produces a vector with a direction given by the right hand rule.

ⁱ McMahon, T., and Bonner, J.T., 1983, "On Size and Life", New York, Scientific American Library

ⁱⁱ Denny, Mark W., 1993, "Air and Water: The Biology and Physics of Life's Media", Princeton University Press

2. Standing up and staying still: forces, Newton's laws, and statics

- 1) What needs to be explained in the motion of things?
- 2) Newton's first law: motion requires no cause
- 3) Newton's second law: how forces change motion
 - i. Quantifying forces
 - ii. Units for force
- 4) Newton's third law: everything is an interaction
 - i. How can this be true? Thought experiments
- 5) Classifying forces
 - i. Contact and non-contact
 - ii. Passive and active

Physics for the Life Sciences: Chapter 2

2.1.0 What needs to be explained about motion?

Life on Earth faces many challenges. One of the most basic is dealing with the constant pull of the Earth's gravity. Every living thing near the surface of the Earth (and every nonliving thing too) is constantly pulled downward. This downward pull is so steady and omnipresent we usually forget it's there. But just one misstep on the staircase, a moment's loss of balance on your bike, one slip of the cup off the edge of the table, and you're reminded of the power of gravity with shocking suddenness. It's not a stretch to claim that gravity is America's number one killer, and it is certainly the number one mechanical challenge for life on land.

In addition to standing up to gravity, many living things have to move themselves and their stuff around. They also need to manipulate things; digging holes, peeling fruit, throwing rocks, taking bites out of their food. To accomplish any of this, living things need to be able to apply forces which they generate. So our first big task will be to understand **forces** and how they affect things. This topic in physics is called "mechanics", one of those otherwise everyday words which means something quite special in physics.

Our understanding of mechanics is based on a few simple principals. They are traditionally summarized in three terse laws first collected by Isaac Newton in the 17th century. Newton's laws provide the tools we need to understand most everything about how objects react to forces. They allow us to predict motions as various as the orbits of the planets and the swimming of a bacterium. Their ability to analyze almost every mechanical situation bigger than a handful of atoms makes them a remarkable part of the collective human intellectual legacy. They're also incredibly useful for understanding what's going on around you, and hopefully by the time you finish studying them you will see the world and what happens in it in a new and richer way.

We will begin by analyzing in detail objects which aren't moving; a subset of mechanics sometimes called "statics". On Earth objects sitting still always experience a number of forces, but these are balanced, so that the total force on them is always zero. Once we have statics in hand, we'll look at cases where the forces are not balanced, and learn to understand they ways in which unbalanced forces cause changes in motion. While the principles of mechanics are simple, there is much to elaborate on. Even the introduction presented here will take up the next seven chapters.

2.2 Newton's first law

Newton's first law is this somewhat surprising assertion:

Any body continues in a state of rest, or of uniform motion in a straight line, unless it is compelled to change its motion by unbalanced forces imposed on it.

This is usually called the "law of inertia". It's not something Newton discovered, as even he would have freely acknowledged; everyone working seriously on motion at the time knew about it. Still, it is hardly obvious. The first bit is no surprise. Objects at rest stay at rest unless you do something to move them. But the second part isn't familiar at all.

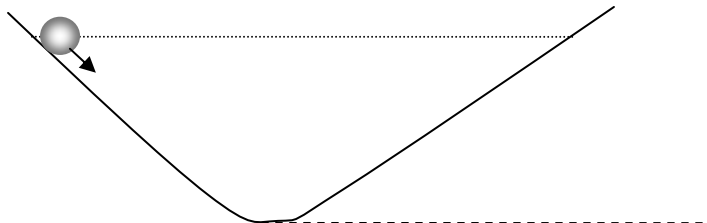
Daily experience does not suggest that an object in motion usually stays in motion. What do you have to do to keep an object in motion? You have to push all the time. Aristotle, noting this ubiquitous experience, assumed that the 'natural state' of an object was to be at rest, and that to have an object in motion required a motive force. He believed that motion implies a mover. But even he allowed a tantalizing exception; objects in free-fall seemed to fall only because it was in their nature to do so. No 'mover' was required to create free-fall motion.

Aristotle gets a bad rap in the teaching of modern science, and perhaps we dismiss Aristotelian beliefs as ridiculous too quickly. In fact Aristotelian views are consistent with much of ordinary experience – objects in motion stop unless someone pushes on them. Seeing behind these "obvious" facts requires great care. It took a world of smart people thousands of years to see what you're learning now. Even with all the facts laid out, you may still find it difficult to see what Aristotle could not.

It was Galileo Galilei, one of the delightful Italians of the 1600s, who first clearly revealed the flaws in the Aristotelian idea of motion. His argument, which elegantly encapsulates the idealization which has proven so powerful in physics, went like this:

1. Imagine a wedge shaped track. Roll a ball down one side and it rises up the other to *almost* the same height
2. Carefully clean and polish the track and the ball rolls still more closely to the starting height, so we might ascribe any remaining loss of height to friction between the ball and the track.
3. Now decrease the angle of the second side. The ball still rises to the same height from which it was launched, but now travels much farther along the ramp.
4. Carry this to its logical conclusion: if we lower the second side until it is horizontal, the ball will travel forever, always attempting to rise again to its original height.

Notice the details here: the ball will roll forever *with no help from anything*. Its 'natural state' is to be in motion, and friction is the only thing which prevents that motion. Without friction anything that was moving would continue forever.



Why was it so difficult for people in the 1800 years between Aristotle and Galileo to understand this? The problem is that friction is acting just about all the time: it is very difficult to see motion without it. This is why, when you look around you, almost everything you see is at rest relative to everything around it. We will talk about friction and how it works quite a bit in this book; it is interesting and extremely important practically. But today we are interested in seeing the world without it. Galileo couldn't quite do this. He couldn't actually make a world without friction. But he could *imagine* it, and it was this imagined idealization that allowed him to recognize the law of inertia.

Where might you have seen motion without friction? The wonderful feeling you get gliding along while ice skating or coasting on your bicycle comes because, unlike most of your motion, this is 'effortless'. Nothing is required to keep you going. When you are almost free from the shackles of friction, you get to experience the law of inertia. A thrown ball or a fired arrow also moves along without obvious influence from friction. Aristotle had some trouble trying to explain this. What 'mover' causes the motion of the arrow?

There aren't many cases of really frictionless motion in nature, but there are a few. Perhaps the most impressive is the motion of the planets, which have continued to circle the sun for billions of years without slowing down or needing to stop to refuel.

The law of inertia is absolutely crucial for understanding motion. Take for example what happens if you smash your car into a brick wall. The front of your car strikes the wall. The force impressed on the car by the wall causes *it* to stop moving forward. But that force is not directly impressed on *you*. Unless something happens to you, you'll just continue to move forward, in a continuous straight line, until something *causes* you to stop. Hopefully this will be your seatbelt or at least your airbag, but if not, it will be the dashboard, the windshield, or the wall. You will not stop, your motion will remain unchanged, until an adequate force is applied to you.

The essential point is that 'rest' is not the natural state of thing. Uniform motion in a straight line is the natural state of things. Objects at rest are just a special case of this. The most important thing the law of inertia provides is a way to tell when unbalanced forces act. If you see an object moving at a constant rate in a straight line, you know there is no unbalanced force acting. If, on the other hand, you see the motion of an object change, then you know for sure that an unbalanced force has acted on it. **The first law tells you how to know whether an unbalanced force has been applied.**

This rule had immediate and important implications for Newton. He knew, for example, that planets travel around the Sun in elliptical orbits; not in straight lines. Applying his first law, this required the action of a force pulling the planets toward the Sun. He knew, of course, that the Earth pulled objects toward it. One of his greatest achievements was to recognize that the same force which pulled an apple toward the Earth also kept the Moon in orbit around the Earth, and the Earth in orbit around the Sun. By examining the motions of the planets, he divined a universal law of gravity, and made robust predictions for the motions of planets, moons, asteroids, and satellites. The key to this was the first law.

2.3 Newton's second law

Newton's second law is essentially of a quantification of the first. The first said that the motion of an object will not change unless an unbalanced force acts on it. The second tells us just how much force is needed to create a particular change in motion. In another sense, the second law is a quantitative definition of a force. If you want to know what force acted, you need only examine how it altered the motion of the object on which it acted.

To precisely state the second law, we need to quantify how much motion an object has. How to might we do this? Is motion a scalar, something with only a magnitude? Or is it a vector, something with both a magnitude and a direction? When an object is moving, it seems natural to care about which way it's moving. So for starters we will look for a vector measure of motion. A bit later we will see that it's sometimes also interesting to just ask whether things are moving at all, without regard for the direction they're going. This would be a scalar measure of motion.

Experience suggests that there's something more to quantifying motion than just measuring how fast things are going. There is a difference between a ping-pong ball and a minivan, even when each is approaching you at the same 20 miles per hour. So in addition to knowing how fast things are going, we'll need to include how much stuff is moving.

With these general ideas in hand, we will define two different measures of motion; a vector measure and a scalar measure.

- **Momentum:** a vector measure of motion. Momentum is often denoted with the symbol \vec{p} , and it is calculated by multiplying the mass of an object by a vector which represents its velocity. Written as an equation we have $\vec{p} = m\vec{v}$. Since mass is a scalar and velocity is a vector, the momentum is always in the same direction as the velocity. The dimensions of momentum are ML/T, and in SI units we would measure it in kilogram*meters / second.
- **Kinetic Energy:** a scalar measure of motion. Kinetic Energy is usually denoted by the symbol KE (which you will note has no little arrow over it; it's just a scalar). The definition for this scalar measure of motion is $KE = 1/2mv^2$. The "v" in this equation is just the magnitude of the velocity vector. The dimensions of kinetic energy are ML^2/T^2 , which in SI units would be kilogram*meters² / second². This combination has a special name. One kgm^2/s^2 is called one Joule (J).

With motion quantified in these ways we are in a position to measure changes, and seek to quantify their causes.

Quantifying Force

Now that we have a way of quantifying motion, we can write Newton's second law formally and quantify force. The usual way to write Newton's second law is:

$$\vec{F}_{total} = \frac{d\vec{p}}{dt}$$

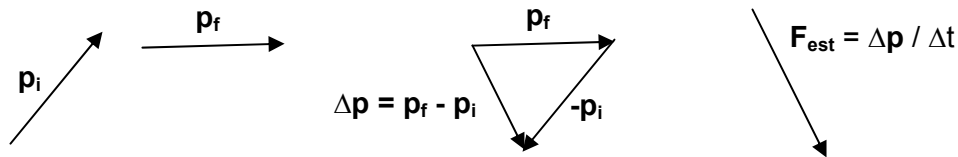
Put into words, the total force acting on an object is equal to the time rate of change of the object's momentum. When you see the momentum of an object changing, you can find the total force which acts by examining how the momentum is changing. One little note about terminology. Many physics texts talk about the “net” force on an object, by which they mean the total, vector sum of all the forces which act. To me, it's much clearer to simply say “total” force, and we'll usually do that.

The dimensions of a force defined in this way are ML/T^2 , which in SI units would be kilogram*meter / second². This combination also has a special name: one kgm/s^2 is called one Newton (N). Many other units of force are used in the scientific and technical literature. All can be converted to Newtons using simple conversion factors, and we will use Newtons exclusively.

Because you're probably not used to thinking about the time rate of change (the derivative) of a vector, it might be useful to consider an example. Imagine the momentum of an object changes by some amount $\Delta\vec{p}$ over some period of time Δt , and we want to estimate how much force was required to create this change. Since the force at each instant is defined by the instantaneous rate of change of the momentum, we can estimate it by dividing the total change in momentum by the total amount of time this change took. Writing this out:

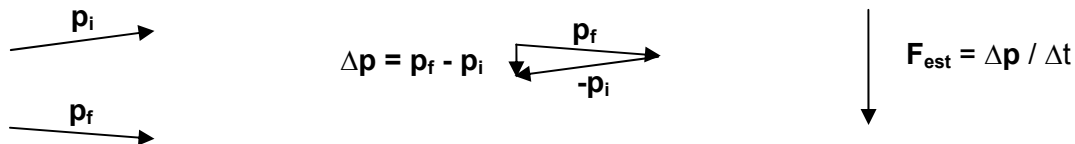
$$\vec{F}_{est} = \frac{\Delta\vec{p}}{\Delta t}$$

Let's look at this with vectors.



Notice that the estimated force vector is in the direction of the *change* in momentum. It is not, in general, in the direction of the momentum itself. Notice too, that although we have drawn the momentum and the force on the same picture, they have different units and hence their lengths cannot be simply compared.

To improve our estimate of the force, we need only consider smaller and smaller time periods, and correspondingly smaller changes in momentum. For example:



Notice that although the change in momentum $\Delta\vec{p}$ is smaller here, the estimated force remains large, because the corresponding time period Δt is also smaller. When we take this notion to its logical limit

$\Delta t \rightarrow 0$ and the force estimated in this way becomes the exact force as defined by Newton's second law above.

It's useful to think about the relation between force and change in momentum in different ways to help see it in all its forms. When momentum changes very suddenly, its time derivative $d\vec{p}/dt$ will be large. Large forces are required to make sudden changes in motion. When momentum changes gradually, its time derivative $d\vec{p}/dt$ will be small. Only relatively small forces are required to make such gradual changes in motion.

If you want to create a certain change in momentum $\Delta\vec{p}$, what do you have to do? For example, imagine that you want to stop an object that begins with a momentum $\vec{p}_{initial}$. The final momentum of the object $\vec{p}_{final} = 0$, so the change in momentum required can be easily calculated:

$$\Delta\vec{p}_{\text{to stop}} = \vec{p}_{final} - \vec{p}_{initial} = -\vec{p}_{initial}$$

You can rearrange our estimate for the force to see what's required to make this happen: $\Delta\vec{p} = \vec{F}\Delta t$. This emphasizes that a particular change in momentum can be achieved with either a large force applied for a short time or a small force applied for a long time. In fact, this quantity $\vec{F}\Delta t$ has a special name; it is called "impulse". So it is sometimes said that you "apply an impulse" to achieve a certain change in momentum.

For the moment, we'll set aside the second law, because we're going to first spend some time analyzing static cases, where the momentum doesn't change. We'll return to dynamic cases, where momentum does change, in Chapter 6.

2.4 Newton's third law

The most important and perhaps least obvious fact about forces is that they are never isolated; they happen **only** in pairs. All forces are two-way interactions. Interactions take place between things; there is never a force that comes from nowhere and pushes on something. Every force comes from one thing and pushes (or pulls) on another.

Newton, realizing this, was the first to recognize a fundamental fact about interactions: they are always perfectly balanced. When one object applies a force to a second, the second applies an exactly equal and opposite force on the first. This is true in every case. Every interaction involves two forces, equal in magnitude and opposite in direction, one acting on each of the two interacting bodies.

This 'third law' of Newton is often stated in the somewhat arcane form Newton used:

For every action there is an equal and opposite reaction.

But it is perhaps more useful to rewrite this in more modern terms as:

When object A exerts a force on object B, object B exerts an equal and opposite force on object A.

Written as an equation:

$$\vec{F}_{A \rightarrow B} = -\vec{F}_{B \rightarrow A}$$

Let's be careful about the notation here. I have written $\vec{F}_{A \rightarrow B}$. By this notation I mean the force applied by object A on object B. Likewise the notation $\vec{F}_{B \rightarrow A}$ denotes the force applied by object B on object A. The minus sign in the way we have written the 3rd law just reflects the fact that while these two vectors are equal in magnitude, they have exactly opposite directions.

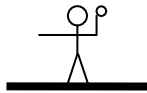
Later in the course we will see that this third law, so surprising at first, has a very deep origin in physics. It is ultimately related to the simple fact that the laws of physics are the same everywhere.

How can the third law be true? Third law thought experiments

The 3rd law is simple to state, but quite surprising. No one before Newton ever recognized it. It almost seems it can't be true, for it seems to suggest that every force which exists will be balanced out in some way. If every time I push on something, it pushes back equally on me, how can anything ever get anywhere? Don't those two forces always cancel out?

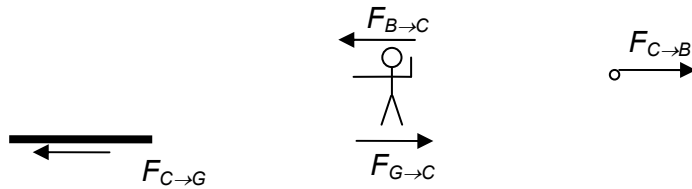
In physics it is often possible to illuminate a principle by considering examples and mentally working through their implications. Application of 'thought experiments' like these led Galileo to a law of inertia, and have always played a central role in physics. A few thought experiments may help to clarify how the third law plays out in the world.

Consider first the way a child throws a ball. To do this, she holds the ball in her hand then pushes it forward with a force $\vec{F}_{C \rightarrow B}$;



The 3rd law tells us that $\vec{F}_{C \rightarrow B} = -\vec{F}_{B \rightarrow C}$: the force the ball exerts on the child is equal and opposite to the force the child exerts on the ball. But the ball goes flying off, while the child does not. What's going on?

This asymmetry of outcome needs explanation. To begin with, the force of the child on the ball is the only force acting on the ball, but the force of the ball on the child is *not* the only force acting on the child. A very simplified view of the situation is drawn below:



Note that I have **not** drawn this picture at a time after the ball was thrown, I have just drawn separate pictures of the ball and the person **while** it is being thrown. After it is released there is no longer any interaction between the ball and the person. To simplify this picture I have also left out the force of gravity which pulls down on both the person and the ball, along with counteracting upward forces that balance the pull of gravity. We'll have much more to say about this a bit later.

The forces on the child come from the ball ($\vec{F}_{B \rightarrow C}$) and from friction between the child's feet and the floor (the force of the ground on the child: $\vec{F}_{G \rightarrow C}$). These two forces both act on the child, and can easily cancel out, leaving the *total* force on the child zero. Her motion doesn't change. Meanwhile the total force on the ball is not zero, so its momentum is suddenly increased. The asymmetry of outcomes for the child and the ball, seen in this light, has an obvious cause. Despite the third law guarantee that the forces they apply on one another are equal and opposite, the full circumstances for the ball and the child are not the same, and hence their outcomes are different.

What if we tried to make this much more balanced, perhaps by having the child throw the ball while standing on a perfectly slippery surface (wet ice or worse) incapable of applying a frictional force? In this case, illustrated below, there would be a net force on both the child and the ball.



While the total forces on each are now exactly the same, the outcomes for the child and the ball remain different. Why?

Newton's second law tells us that a force acting for some period of time produces a change in momentum, $\Delta \vec{p} = \vec{F} \Delta t$. Since a force of the same magnitude acts for the same amount of time on both the child and the ball, it must produce the same change in momentum in each. Because the mass of the ball is relatively small, it must have a large change in velocity to create this change in momentum. Because the mass of the child is relatively large, she will experience only a small change in velocity to create this change in momentum. The same force acting on different objects can produce disparate outcomes even in very simple cases.

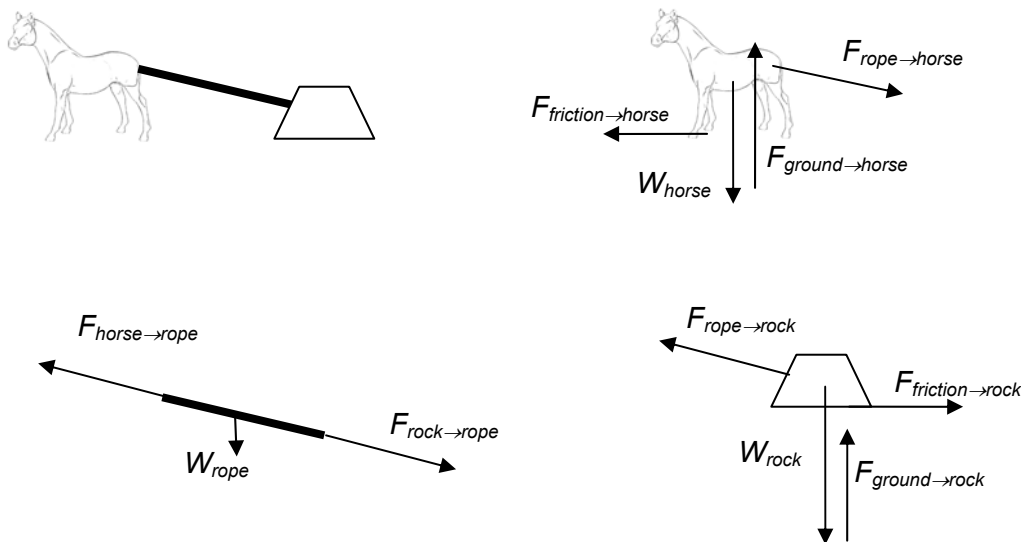
Analyzing forces: the free body diagram

The analysis above provides a first example of a technique we will find essential in understanding mechanics; the construction of a ‘free body diagram’ for each element of a problem. Newton’s first two laws tell us that the motion of an object is completely determined by the total force which acts upon it. This is Newton’s most essential lesson: to understand change, focus on the forces. Newton’s third law gives us an important clue about these forces, telling us that they always occur in equal and opposite pairs.

To put these ideas to use, you take an example like the girl throwing a ball above and begin by taking it apart. The first step is to draw each object in the problem alone, separated on the page from all the others. Then begin to identify the forces which act on each object, always remembering that every force is part of a ‘third law pair’. This picture, with every object you care about drawn separately, and with all the forces which act on them identified, is a free body diagram. It is the essential first step in the analysis of a mechanics problem. To illustrate how this approach works, let’s consider a slightly more complicated case, one with three objects to keep track of.

Consider what happens when a horse drags a stone forward at a constant speed with a rope. This situation has three elements: the horse, the rope, and the stone. We begin by drawing each of these three separately. Then we identify the forces which act on each object.

For the horse, there are four. The first is its weight, the downward pull which the Earth’s gravity exerts. Then there is an upward force exerted on the horse by the ground beneath its feet. It is this force which prevents the horse from plummeting to the center of the Earth. To pull forward, the horse plants its feet on the ground and pushes backward. When the horse pushes backward on the ground, the ground pushes forward on the horse. Finally, the horse pulls forward on the rope, which then pulls backward on the horse.



For the rope, there are three forces to keep track of. The first is the force with which the horse pulls on the rope. This is part of the third law interaction between the horse and the rope. The horse pulls on the rope, and the rope pulls on the horse. A second force on the rope is the force with which the rock pulls on the rope. This is part of the third law interaction between the rope and the rock. The rope pulls on the rock, and the rock pulls on the rope. The third force on the rope is its weight.

Finally we have the forces on the rock. They mirror the forces on the horse: the forward force of the rope on the rock, the backward force of friction with the ground, the downward force of the rock's weight, and the upward force of the ground resisting this weight.

Much can be learned from these three pictures, without doing any calculations. First, all three objects are moving forward at constant speed. Since their momenta do not change, the total force on each must be zero. These force sums are vector sums, and since they sum to zero, the sum of the components along each direction (vertical and horizontal for example) must be zero.

For the horse; notice that the force of the rope on the horse points partly downward. Both the downward component of this force and the weight of the horse must be counteracted by the upward force of the ground on the horse. This tells us that this upward force must be larger than the weight, something which would not be true if the horse was just standing there. Similarly the force of the rope on the rock is partly upward. This suggests that the upward force of the ground on the rock is somewhat reduced relative to the rock's weight.

The rope is mostly just stretched forward and back, pulled forward by the horse and backward by the rock. In the process the rope is stretched. It will do whatever it can to "ease" the stretch, so it pulls the rock forward while also pulling backward on the horse. These two forces are almost, but not quite identical. You can see this from the free body diagram. Since there is a (presumably small) downward weight of the rope, the upward part of the force of the horse on the rope must be somewhat larger than the downward part of the force of the rock on the rope. This makes the two forces on the ends almost, but not quite, equal to one another. In a sense the rope just "transmits" a force between the horse and the rock. This transmission of force is not quite perfect because of the weight of the rope. For a massless rope, the transmission of force would be perfect.

When faced with a mechanical situation of any kind, the essential first step will be to consider the forces which act on each object in the problem. Creating a correct free body diagram for each is the most important step toward a precise analysis.

2.5 Types of forces

Newton's three laws are adequate to allow us to understand a truly extraordinary range of phenomena, including everything to do with the structures of living things and how they move around. At a basic level Newton's laws tell us to understand motion by paying attention to forces. To do this, we're going to spend the next several chapters pondering different kinds of forces and learning how they act. We begin here by classifying forces in two very basic ways.

Every force can be put into one or the other category in each of three different general schemes. The first way to classify forces is to ask whether they are **active** or **passive** forces.

Active forces are those whose magnitudes are determined by some external factor. Good examples are pushes and pulls (where an active participant like a person decides how large the forces will be), the gravitational force (which is always the same for a body near the Earth's surface), and electric and magnetic forces we will learn about later in this course.

Passive forces are different from these. Passive forces are those which arise, and adjust themselves, in *response* to active ones. A good example emerges when you push on a wall. The wall pushes back on you just enough to counteract your active push. If you lean gently against the wall, it pushes only a little. If you race at the wall and smash into it, it exerts a large force back on you. The key point with passive forces is they are *whatever they have to be*, and what they have to be is determined by the active forces which are present. There are limits to passive forces. If you run into the wall hard enough, it will not be able to create a large enough force to stop you. Instead the wall will apply the biggest force it can and then break.

There is a second, independent, way to distinguish between forces; we can classify them as **contact** and **non-contact** forces.

Contact forces are those that arise from intimate physical contact between two bodies. They act only when the bodies touch, in the usual sense of having their surfaces approach one another at atomic scales. Examples of these very common contact forces are those we associate with pushing against a wall or placing a book on the table, the forces which occur in a collision, and the frictional forces which allow you to walk across the room.

Non-contact forces are forces which can act even when the two bodies are not touching one another. For this reason they're also called long-range forces. The only non-contact force we'll talk about extensively in this class is gravity. You don't have to be touching the ground for the force of gravity to act on you. Just step off a chair and you will see what I mean. In fact gravity can act on you even when you are very far from the Earth, through completely empty space. There are other non-contact forces. One which you have probably seen a little is the magnetic force, which like gravity can act even through empty space.

A third useful way to divide forces into categories is to talk about **fundamental** and **phenomenological** forces.

Fundamental forces are those which we understand relatively deeply, like gravity. The laws describing these forces seem to reflect an underlying reality at a level which suggests that they are "True" with a capital T. Their behavior tends to become simpler and simpler as we look at them more closely. It turns out there are only four of these fundamental forces known in nature, and even some of these are closely coupled.

The four fundamental forces are:

- Gravity: Every object with mass attracts every other. This attraction holds planets, stars and galaxies together, and keeps you on the surface of the Earth.
- Electricity & Magnetism: This combination is responsible for chemistry, and all the bonding between atoms that makes matter interesting. Every force you'll see in this class, *except gravity*, ultimately arises from electricity and magnetism.

- Strong Nuclear Force: This very short range force holds atomic nuclei together, allowing for all the existence of the various elements of the periodic table.
- Weak Nuclear Force: This short range force is responsible for the radioactive decay of some atomic nuclei.

While both the strong and weak force are crucial for the existence and nature of chemical elements, they are also remote, in the sense that about all they do is help create the periodic table. You won't see them acting more directly in your lives.

Phenomenological forces, unlike fundamental forces, are inescapably complex, like the force of friction. While all forces ultimately arise due to the four fundamental forces, this is often far from clear. When forces are more complex, we describe them with "laws" that would more appropriately be called models. In them we attempt to quantify an often very complicated set of phenomena by a series of approximations. The distinguishing feature of phenomenological models is that the more closely you study the phenomena, the more complicated the law you must use to describe it becomes. This is considered evidence that the understanding you have is ad-hoc, approximate, and not fundamental.

That doesn't mean phenomenological models for forces are not accurate reflections of reality or that they aren't "true". It's just that by acknowledging that they gloss over details, we confine them to being "true" with a lower-case t. We know for sure that there are other details hiding beneath these general principles.

We'll recall these three divisions a little as we move through our discussion of mechanics. Keeping these ideas in mind will help you to know how different forces act.

A Quick Summary of Some Important Relations

Newton's Laws:

1. Unbalanced forces alter motion, but are not required to maintain motion
2. The size of a force is measured by how rapidly it changes momentum
3. Every force is part of an interaction between two objects

$$\vec{F} = \frac{d\vec{p}}{dt}$$

$$\vec{F}_{\text{est}} = \frac{\Delta\vec{p}}{\Delta t}$$

$$\vec{F}_{\text{A on B}} = -\vec{F}_{\text{B on A}}$$

3. Forces and structures

- 1) An active, non-contact example: gravity
 - i. Gravitational attraction: a property of mass
 - ii. Newton's universal law
 - iii. An approximation important for life: earthly weight
- 2) A richer example: forces in the contact of solids
 - i. A first component: the normal force
 - ii. More on free body diagrams
 - iii. Examples of determining the normal force
 - iv. The second component: friction
- 3) Transmitting forces: ropes and tension
 - i. Forces within a rope
 - ii. Ropes and sending forces around corners
 - iii. Why massless ropes?
 - iv. Tension and force transmission
- 4) Quantifying the ability of a force to cause rotation
 - i. Two ways of thinking about torque
 - ii. Torque and the vector product

Physics for the Life Sciences: Chapter 3

In this chapter we will start exploring the application of Newton's laws. We begin by learning about a few forces and how to model their action. For now, we will focus on gravity and a basic model of the contact forces between solids. With this model of forces in place, we have enough to analyze a wide variety of structures, from elephants to the Parthenon.

3.1 A first force: gravity

The first specific force we will talk about is gravity, the perfect example of an active, non-contact force. It is an active force because its strength is determined by a very specific law – Newton's 'universal' law of gravitation. The gravitational force on an object doesn't depend at all on what other forces act. It is a non-contact force because it very freely acts at a distance, even a very great distance, and even through empty space.

Gravity is a fundamental force. Every object with mass attracts every other object with mass through this force, though as we will see, the gravitational force between two everyday objects is usually very weak. It only becomes large enough to notice when at least one of the objects involved is *very* massive, something like the Earth. The gravitational force exerted on things near the Earth's surface can be modeled very simply, as described in the next section. After introducing this simple, limited model, we will examine a more fundamental model for gravity: Newton's universal law of gravitation. Even Newton's very accurate model of gravity is not perfect, and we will finish our introduction to gravity with a few words about our best current model for its action, Einstein's general theory of relativity.

Weight: an active, non-contact force

All objects near the surface of the Earth are pulled toward it by its gravitational interaction. Physicists call this downward attraction of the Earth an object's 'weight'. For objects close to the surface of the Earth, there is a very accurate (though approximate and phenomenological) force law which predicts the magnitude of this attraction:

$$W = F_{\text{Earth-Object}} = m_{\text{Object}}g$$

In this equation, m_{Object} is the mass of the object in question, and g is a constant which has a value

$$g = 9.8 \text{ m/s}^2$$

Multiplying the mass of an object in kilograms times this constant g gives the magnitude of the downward force on the object in Newtons. All objects near the surface of the Earth experience this force, no matter what other forces act. So this is the one force which is always present on every living thing we know about. The free body diagrams you draw for mechanical analysis of anything near the surface of the Earth will include this force for each object.

There are several details to consider carefully when you think about weight. The first has to do with Newton's third law. The third law says that if the Earth is pulling you down with a force W , you must be pulling the Earth upward with a force which is equal and opposite. That is:

$$\vec{F}_{\text{You-Earth}} = -\vec{F}_{\text{Earth-You}} = -\vec{W}_{\text{You}}$$

While the Earth pulls you down, you pull it up. This action-reaction pair is always there.

If this is so, why doesn't the Earth come rushing up to meet you when you jump off a cliff? After all, you're pulling the Earth upward just as hard as it's pulling you down. The reason for this disparity, which always emerges in unequal interactions like this, is the relatively enormity of the Earth's mass compared to yours. Your weight ($\vec{F}_{\text{Earth-You}}$) is a big enough force to quite easily change your motion. But a force of the same size ($\vec{F}_{\text{You-Earth}}$) is much too small to make an appreciable change in the motion of the Earth. So although these two forces act to pull you and the Earth together, it's you who does all of the moving.

A second point worth discussing in some detail is the sensation of weight. What is it we feel as our own weight? Can you feel the force of gravity upon you? Imagine what happens when you jump off a chair. For a moment you are floating freely in the air. Do you "feel" a force tugging on you when you do this? While you're in the air, there is no sensation of force at all. Try this if you dare, but please be careful. Don't jump off anything higher than a chair! So what is the sensation of weight? What is this feeling you get when standing around all day?

When you are standing on the floor, the sensation that you feel is the upward pressure of the floor on your feet. Now think about what you feel when you sit in your chair. Is there any pressure on your feet? Now the pressure seems to be on your backside. And if you stand on your head you 'feel your weight' on your

hands and head. So could what we feel really be the weight? Is what you feel really the force exerted on you by the gravity of the Earth? In fact it is not. The sensation you feel when you talk about weight is actually the force which something else applies to you to resist the downward pull of gravity, to prevent you from falling downward.

When you are standing still your weight (the force of gravity on you) pulls you down. To remain stationary, some other force must balance this. That balancing force is provided by an object you are in contact with. The sensation you feel as weight is just the force of the floor (or your chair, or whatever) pushing *up* on you to resist your weight. Take away the floor, or the chair, and your sensation of weight would vanish, but the weight itself, the downward pull of gravity, would not.

Newton's universal law of gravity and the origin of 'g'

Gravity is one of the four fundamental forces of physics. It can be described very accurately in a remarkably simple law originally proposed by Newton. He used it to explain both gravity on the surface of the Earth and the motions of the planets in the solar system. This was a tremendous surprise in Newton's time, when the workings of the heavens were widely thought to be completely unrelated to what happens on Earth. For this reason, Newton's formula is called the 'universal' law of gravitation. Today we have strong evidence that *every* law of nature is universal; the same laws of physics which act here on Earth apply absolutely everywhere in the cosmos. Despite this, we still honor Newton's achievement by singling out this one law for its universal nature.

Newton's law describes a 'central force', which always pulls two objects directly toward one another. The magnitude of the gravitational force between objects A and B can be written:

$$F_{AB} = \frac{Gm_A m_B}{r_{AB}^2}$$

Where m_A and m_B are the masses of the two objects, r_{AB} is the distance between the two objects, and G is a universal constant which sets the scale for the strength of the gravitational force. The value of G is about $6.67 \times 10^{-11} \text{ N(m/kg)}^2$.

Gravity, perhaps surprisingly, is a very weak force. You can see this by applying Newton's equation to an ordinary situation. Two one kilogram objects placed one meter apart attract one another with a force of $6.67 \times 10^{-11} \text{ N}$. You might compare this force to the weight of an apple, which is typically about 1 N; ten billion times larger.

If all the objects we encountered were about our size, we might never notice gravity. But one thing near us is much, much bigger: the Earth. The downward gravitational force the Earth exerts on us, our weight, is the most familiar example of gravity. How can we use Newton's universal law to understand the gravitational attraction between a person and the Earth? After all, some parts of the Earth are very close to us, while others are far away. Each part exerts its own small force on us. To find the total we would have to construct the vector sum of millions of tiny forces.

Fortunately, the precise nature of the gravitational force law allows a remarkable simplification. For a force of just this form (dependent on the product of the masses and inversely proportional to the square of the distance) it can be proven that the force exerted by a spherical shell of mass on an object outside the shell is exactly the same as the force which *would be* exerted if all of the mass of the shell were located at its center. This ‘shell theorem’, first proven by Newton to solve exactly this problem, makes it easy to use the universal law of gravitation to predict the weight of a person standing on the Earth’s surface.

Think of the Earth as constructed of many shells, one nestled inside the next. Each shell exerts a gravitational force exactly as it would exert if all its mass was concentrated at the center. Since each shell does this, the force from the whole spherical Earth is just what it would be if all the mass of the Earth were concentrated at its center. Since a person standing on the surface of the Earth is one Earth radius from this center, the gravitational force experienced by such a person is:

$$F_{\text{Earth-Person}} = \frac{Gm_{\text{Person}}m_{\text{Earth}}}{R_{\text{Earth}}^2} = m_{\text{Person}} \left(\frac{Gm_{\text{Earth}}}{R_{\text{Earth}}^2} \right)$$

Notice that while the mass each person on the Earth is different, the mass and radius of the Earth don’t change. Combining the constants shown in the parentheses above we get:

$$\frac{(6.67 \times 10^{-11} \text{ N(m/kg)}^2)(6 \times 10^{24} \text{ kg})}{(6.4 \times 10^6 \text{ m})^2} = 9.8 \text{ m/s}^2$$

This combination of constants, which has units of m/s^2 , is just the constant ‘g’ invoked in the section above. This derivation shows how the very simple phenomenological force law described above, $W = mg$, emerges from a much more fundamental understanding of how gravity works. It also allows us to consider how accurate we might expect this simple model to be.

First, how accurate is the approximation that everything near the surface of the Earth is actually a distance R_{Earth} from its center? The Earth is not a perfect sphere. It is covered with mountains and deep ocean trenches. These are not, however, very large compared to the size of the Earth. The difference in height from the deepest ocean trench to the top of highest mountain is 19,700 m. This is about 0.3% of the Earth’s radius. An object moving from the bottom of the Marianas Trench to the top of Mount Everest would experience a change in weight of about 0.6%. So this is only important if we are demanding rather high precision.

In addition, the Earth is not a perfect sphere. It is slightly flattened, so that a point on the equator is farther from the center of than a point at the poles. This causes a similarly slight increase in weight, about 0.5%, as an object moves from the equator to the poles. There are other small effects which alter the *apparent* weight, but do not affect the actual gravitational attraction of the Earth, like the buoyancy of the air and the rotation of the Earth. The upshot of this discussion is that none of these effects are important at even the 1% level.

The assumption that $W = mg$ is a very useful approximation for everything near the surface of the Earth. Since this is where life spends its time, this approximation will be fine for most purposes in this class. It will fail only when we consider the gravitational interactions of objects separated by much greater distances, like the Earth and Moon or the Sun and Jupiter.

A consequence of gravity: fluids and the buoyant force

Life exists immersed in fluids; either air or water. Both are fluids because they flow quite freely. They are unable to resist the sideways pull of gravity and hence slip downward until their surfaces become level and they can go no lower. When an object is immersed in such a fluid, gravity will pull down on both the object and the fluid around it. If the object is more dense than the fluid, as is usually the case in air, the object will be pulled downward more strongly than the fluid, and will sink through it until supported by some other means, often the ground. If the object is less dense than the fluid, as is sometimes the case with water, the fluid will be pulled down more strongly, will flow under the object, and provide at least enough support to hold up the object against the pull of gravity.

We will set aside a detailed understanding of this phenomenon for the moment, but note for now its most basic result. Any object immersed in a fluid here on Earth will experience a buoyant force equal in magnitude to the weight of the fluid which the object displaces. We can calculate this buoyant force by multiplying the volume of the object by the density of the fluid and the gravitational constant g .

$$F_{\text{buoyant}} = \rho_{\text{fluid}} V_{\text{object}} g$$

Like the weight, this buoyant force is ever present for objects on Earth, and hence for all life. It is an active force, in this case present only when there is contact between the fluid and the object. It is often useful to compare the magnitude of this buoyant force to the weight of an object, which can be written in a parallel way:

$$F_{\text{weight}} = \rho_{\text{object}} V_{\text{object}} g$$

From this equation it should be clear that the relative importance of these two forces, one pulling down and one pushing up, depends on comparing the density of the fluid to the density of the object. When the density of the object is larger than that of the fluid, weight will be larger, and the object will tend to sink through the fluid until it is supported by something else. When the density of the fluid is larger than that of the object, the buoyant force will be larger, and the object will float upward through the fluid until either its density is matched by the fluid density or it emerges from the surface of the fluid.

The media of life, air and water, have densities which differ by about a factor of 1000, making the importance of buoyancy on land and under water very different. Most living things are largely made of water, and hence have densities within a factor of a few of water.

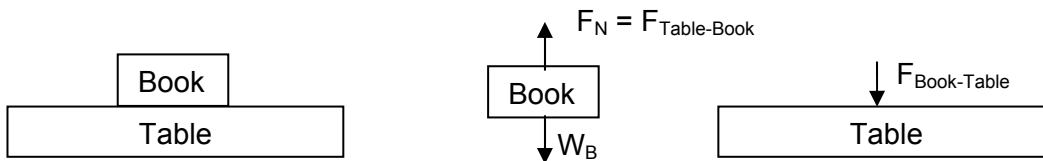
$$\begin{aligned}\rho_{\text{air at sea level}} &\approx 1.2 \text{ kg/m}^3 \\ \rho_{\text{fresh water}} &\approx 1000 \text{ kg/m}^3 \\ \rho_{\text{organisms}} &\sim 500 - 2000 \text{ kg/m}^3\end{aligned}$$

From these numbers it should be clear that, when considering living things, we can largely ignore the action of a buoyant force in air. It plays a negligible role in supporting the weight of an organism. But if that organism lives in water, the buoyant force is certain to play a much more important role. It will always support at least some large fraction of the organism's weight. The support of weight by buoyancy is one of the principal reasons for the very different structures seen in organisms on land and in the sea. We will return to it a number of times in this text.

3.2 The normal force: a passive, contact force

Most of the forces we encounter in our lives are contact forces. They arise from the direct, "touching" interaction between two bodies. When two more or less solid bodies are in such direct, atom-to-atom contact, it will be useful to describe the force between them as composed of two parts; a component perpendicular to the plane of contact between the surfaces (the normal force), and a component parallel to plane of contact between the surfaces (the friction force).

The first of these is the part of the contact force which prevents one object from moving through the other. Let's look at an example.



When I put a book on a table, the book's weight tries to pull it downward. To move downward the book would have to pass through the table. The table prevents this by pressing back up on the book with a force that keeps it in place. This force (which acts to prevent the objects from passing through one another) always acts perpendicular to the plane of contact of the objects. Because of this we will call it the "normal" force. It is NOT normal in the sense of "usual", but normal in the mathematical sense of "perpendicular" to the surface between the two objects.

There's a real mystery here. How can an inanimate object like a table exert a force? More important, how can a table "decide" how much force to apply; pushing up very lightly on a pencil while providing a much larger force to support a lamp? The answer lies in the passive nature of the normal force: it exists only in response to some other force.

Begin by thinking about a cushioned chair. When you sit on it, the chair compresses to some point. To compress it further would require a force larger than your weight. In other words, as it is squashed, it pushes back up on you, harder and harder, until it is pushing up on you with a force equal to your weight. This squashing, this distortion, is what allows the chair to push back up on you. When you push atoms closer together, they push back.

Now imagine something harder than the cushioned chair. If you sit on a plastic chair, it too is distorted until it pushes back on you just enough to balance your weight. Take this to its conclusion; when you stand on the floor, the floor actually distorts until it pushes back on you with a force just large enough to

prevent you from falling through it. This distortion, which always accompanies forces applied by solid objects, is perfectly real, even when it is not apparent.

So this normal force prevents objects from passing through one another. How big is this force; what is its magnitude? The basic answer is "whatever it has to be". For this reason, the normal force is our first example of a passive force. Passive forces have magnitudes which are not determined in advance. They arise, and adjust themselves, in response to active forces. Their magnitudes are determined based on the restrictions which give rise to them. They can be any size from zero up to some limit at which the object creating them breaks.

So, if I push down on a table, it pushes back up on my hand with a force just equal and opposite to my own. If I push harder the resisting normal force increases. If I stop pushing it goes away. The force adjusts itself to be just as large as it needs to be to prevent my hand from moving through the table.

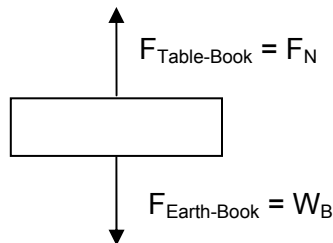
What's the underlying origin of the contact force between solids? Ultimately all of the forces we will talk about in this course (except gravity) are consequences of electromagnetic forces between atoms. The details of these electromagnetic interactions determine whether atoms placed close together will resist being pushed closer together or will attract one another (perhaps bonding and sticking together). Later in this book you will learn quite a lot about the nature of these electromagnetic forces. For now, we will encode a lot of complicated atomic interactions in a few simple phenomenological rules.

More on free body diagrams

Notice what we did there. In order to understand what happens to two bodies *while they interact*, we have drawn each of the bodies separate from the others, so that we can understand fully the forces *on each one*. Let's look at a couple of other examples of this. What happens when a book sits on the table? What are the forces on it: First we might draw the circumstance:



Now, in order to understand it, I draw a free body diagram for *each* part of the problem. First consider the book. What are the forces acting on it? It experiences a weight, the gravitational force of the Earth pulling it downward. Since it is sitting still, it must also experience some other force which balances this. This is the force with which the table pushes back up on the book. This is called the normal force.



Now in order for it not to move we know that all the forces acting on it must balance, so that the normal force and weight must be equal in magnitude:

$$F_N = W_B = m_B g$$

Since the normal force must be opposite in direction to the weight, we know that:

$$\vec{F}_N = m_B g \hat{y}$$

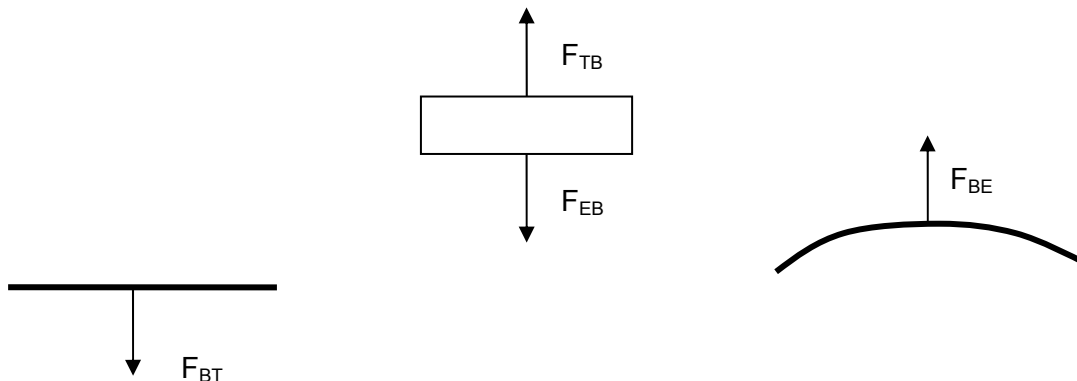
In this equation we note the direction of the normal force with the upward \hat{y} unit vector. What does Newton's third law say about this? It says that for every force there must be an equal and opposite 3rd law partner. What are the third law partners for the forces which act on the book?

The normal force is the force with which the table pushes on the book. The table pushes on the book with a force: $\vec{F}_{TB} = \vec{F}_N = m_B g \hat{y}$, so the book must push on the table with a force $\vec{F}_{BT} = -\vec{F}_N = -m_B g \hat{y}$. These two forces, the force of the table on the book and the force of the book on the table, are equal and opposite.

What about the weight? The Earth pulls on the book with a force $\vec{F}_W = -m_B g \hat{y}$, so the book must equally pull on the Earth with a force $\vec{F}_{BE} = m_B g \hat{y}$. So to understand the (non)motion of the book, we have to consider two different third law pairs:

$$\vec{F}_{Table-Book} = -\vec{F}_{Book-Table} \text{ (Normal force) and } \vec{F}_{Earth-Book} = -\vec{F}_{Book-Earth} \text{ (Weight, or gravity)}$$

And to draw all these forces and keep track of both interactions I would have to draw *three* objects:



This is not really all there is. At this point, we have not drawn *all* the forces acting on either the table or the Earth. The only body I have completely analyzed here is the book. The table must have other forces acting on it, or it would accelerate downward, away from the book. Likewise, the Earth must have other forces acting on it or it would accelerate (however slowly) up towards the book.

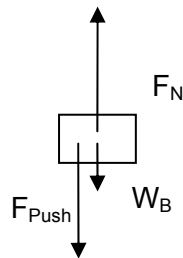
It is very important that you should understand the third law, and be able to identify third law pairs in any situation correctly. If you can't identify third law pairs with confidence it will be very difficult for you to correctly analyze even slightly complex systems.

Examples of determining the normal force

The simplest case to examine is the one we've just done. When I set a book down on a table, the table distorts slightly, bending just until the upward force it exerts perfectly balances the downward force which gravity exerts on the book. So in this case

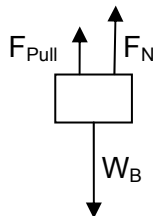
$$F_N = W_B$$

Now imagine that instead of just laying the book on the table, I push down on it with some force F_{Push} . What is the normal force now?



Since the book remains in place, I know that $F_N - W_B - F_{Push} = 0$, or $F_N = W_B + F_{Push}$. How does the table manage to apply a larger force now than it did when I just laid the book on the table? How does it adjust its force to just the right value? This happens because the resisting force the table applies increases as it distorts further. When you push down on the book, the table distorts more, and hence is able to apply a stronger resisting force. It just continues to distort until the force it is applying is just big enough to balance the book.

What if, instead, I pull up somewhat on the book?



Now I know that:

$$F_{Pull} + F_N = W_B \quad \text{or} \quad F_N = W_B - F_{Pull}$$

Notice what happens though, as I gradually increase F_{pull} , F_N continually decreases until it becomes zero. What happens after this? In most cases, the normal force cannot become negative. Typical solids do not stick together when you try to lift one of them off. But of course they sometimes do, as when two sticky surfaces come into contact. This should remind you that when we talk about the normal force in this way, we're talking about an approximate model for a very complex variety of interactions. We have seen simple cases in which the normal force F_N for an object on a table can be either larger **or** smaller than the weight of the object; it may even be zero. In fact the special case in which $F_N = W_B$ is only true in particular cases, and nothing like a general rule.

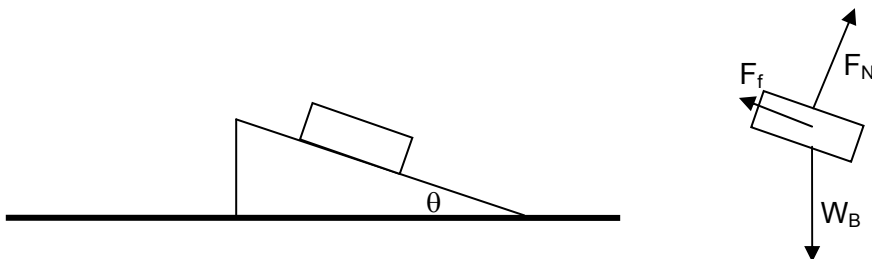
The weight and the normal force are **not** third law partners. The third law partner for the upward force F_N is the downward force of the book on the table, and the third law partner for the downward force W_B is the upward gravitational force of the book on the Earth. Since F_N and W_B are not third law partners, there is no reason that they must be the same. Sometimes they are, but they certainly don't have to be. To determine the normal force F_N in a particular problem, you just have to figure out how large a force is required to keep the objects from moving through one another. You can do this by summing the components of all forces in the vertical direction.

In fact there's another, really simple and obvious way to tell that F_N and W_B cannot be third law partners in this problem: *both forces act on the same object!* Remember, the third law is about forces which are exchanged between two objects. The two parts of a third law pair can **never** act on the same object.

The other part of the contact force: friction

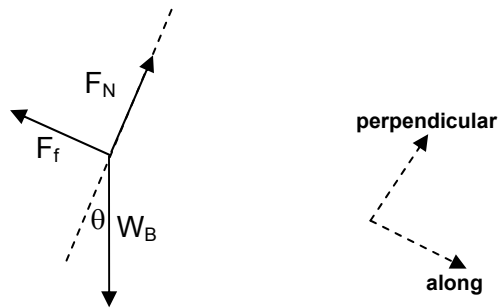
The normal force is what prevents objects from passing through one another. It is the part of the total force between two surfaces which acts perpendicular to their plane of contact. The rest of the force between two bodies is the part which acts parallel to the plane of contact. This force acts to prevent one object from slipping over the other. It resists their relative motion. We call this part of the interaction between two objects the force of friction. We will have a lot more to say about friction and how it works in the next chapter. For now let's just concentrate on the manner in which friction acts. Friction acts in an attempt to prevent relative motion *along* the plane of contact between two objects.

Let's look at a simple example; an object sitting at rest on a slope:



What are the forces on this? There is a weight acting straight down. Then there is some interaction between the book and the surface of the wedge it sits on. This total contact interaction has two parts: a normal force perpendicular to the surface, and a frictional force along the surface. I know (because its motion doesn't change), that the total force acting on it must sum to zero.

$$\vec{F}_{Total} = \vec{F}_N + \vec{W}_B + \vec{F}_f = 0$$



Very often in a problem like this it is useful to work in a coordinate system which defines directions along and perpendicular to the surface between two bodies. Such a coordinate system is shown above. Now we know that since it doesn't move the sum of the forces must be zero, and that in turn means that the sum of the forces in each direction must be zero. Now let's add up these forces in each direction:

$$\Sigma F_{Along} = W_B \sin(\theta) - F_{friction} = 0$$

$$\Sigma F_{Perpendicular} = F_N - W_B \cos(\theta) = 0$$

So in this case we know that:

$$F_{friction} = W_B \sin(\theta)$$

$$F_N = W_B \cos(\theta)$$

Once again we see that the normal force is not equal to the weight (it rarely is), and in addition we see how we can use the lack of motion to figure out how large this "frictional force" must be.

This picture of a block on an 'inclined plane' is the very icon of traditional introductory physics courses. Unlucky students have been learning to analyze these for literally hundreds of years, and they've always seemed completely unconnected from everyday experience. After all, most of you stopped playing with blocks some time ago. But in fact this example, while drawn in an abstract way, is very much an everyday experience. Here are two examples.

The first is standing on a slope: every time you stand on a slope a situation very like what we just described happens. If there is not enough frictional force preventing you from sliding down the slope you

will slip downward. No doubt this is something you're quite aware of, and once snow and ice arrive you'll be careful about standing on slopes.

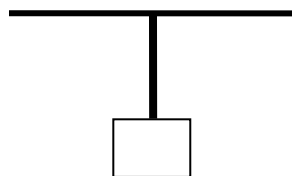


The second example is the slope itself! Each layer of a hill would slip downward if not held in place by a frictional force. Sand dunes, like Sleeping Bear dune in Northwest Michigan, provide a vivid example of this. When they become too steep, the force required to hold the sand in place becomes larger than the frictional force available, and top layers of sand begin to slide downward. The maximum angle for a pile of sand is often given the poetic name 'the angle of repose'. More impressive and dangerous versions of this are seen all the time in avalanches and landslides.

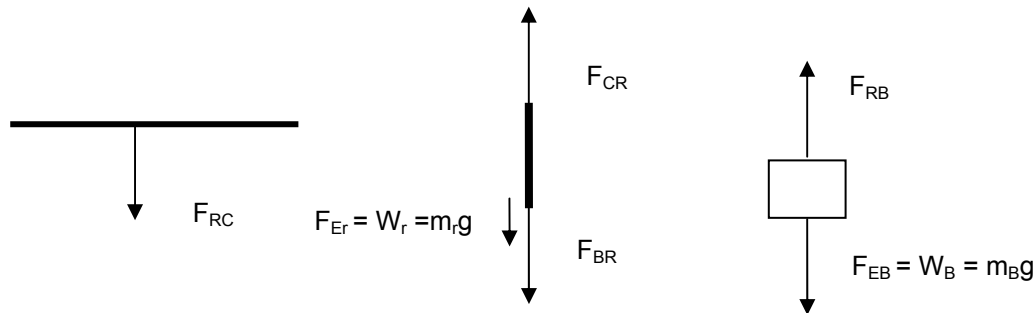


3.0 A way to transmit force; ropes and tension

We have seen how simple contact forces can occur due to the compression of bodies, as when a book sits on a table. It is also possible for objects to exert forces when they are "stretched". This process of attempting to stretch a body is called putting it "in tension". A simple example of how an object in tension behaves is given by a block hanging from the ceiling on a rope.



We can draw three free body diagrams for this: one for the ceiling, one for the rope, and one for the block.



What do we know about these? We know that for the mass at the bottom:

$$\Sigma F_y = F_{Rope-Block} - W_{Block} = 0 \quad \text{or} \quad F_{Rope-Block} = W_{Block}$$

and for the rope:

$$\Sigma F_y = F_{Ceiling-Rope} - W_{Rope} - F_{Block-Rope} = 0 \quad \text{or} \quad F_{Ceiling-Rope} = W_{Rope} + F_{Block-Rope}$$

We also know that some of these are third law partners, so their magnitudes must be equal. In particular:

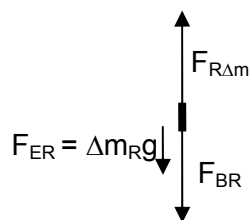
$$F_{Rope-Ceiling} = F_{Ceiling-Rope} \quad \text{and} \quad F_{Rope-Block} = F_{Block-Rope}$$

So now we can say:

$$F_{Ceiling-Rope} = W_{Rope} + F_{Block-Rope} = W_{Rope} + W_{Block}$$

So the force which the ceiling must exert to support the rope and the weight is just equal to the weight of the mass plus the weight of the rope. Not too surprising.

Now let's think about what's really happening *inside* the rope. Picture the little piece of the rope at the bottom. The mass is pulling down on it with a force W_B . This little piece has some mass $\Delta m_r g$, so we can write the same kind of free body diagram for it:



So for this little piece we find:

$$\Sigma F_y = F_{Rope-\Delta m} - W_{\Delta m} - W_{Block} = 0 \quad \text{or} \quad F_{Rope-\Delta m} = W_{\Delta m} + W_{Block}$$

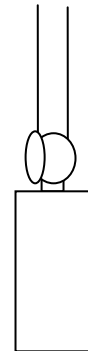
As we gradually move up the rope, this force inside the rope gradually grows, until just at the top, where it must support the full weight of the rope below it, the internal stretching force in the rope is supporting the full weight of both the entire rope and the block.

What's happening is that each little part of the rope is being pulled equally both up and down; these forces are trying to tear the rope apart. It is the ability of the rope to hold itself together against this “tension” that allows it to transmit the weight of the hanging mass to the ceiling above.

It is also possible to transmit forces with compression. That's what I do when I push something along with a stick. And there's no need for the thing in tension to be a rope. I would just as well hang the mass from typical solid like a wooden meter stick and the analysis would be exactly the same. There would be a tension in the stick.

Most solids can support loads either in tension or in compression. Some others, especially those with more interesting internal structure like ropes, tendons, and flesh, are better at supporting loads in tension than compression. As we will see, your body is basically a framework of bones capable of supporting your weight in compression, with flesh and organs hanging from this skeleton. Most of these other parts are held in place by materials stretched in tension.

Let's apply this to a real world example. You're headed to the airport with a suitcase to check. It has the maximum allowable weight (50 lbs, or about 23 kilograms), and you carry it with your arm straight up and down. The picture is to the right. The analysis is the same as above:



$$\begin{aligned} F_{\text{shoulder-arm}} &= F_{\text{arm-shoulder}} = W_{\text{arm}} + W_{\text{suitcase}} \\ F_{\text{arm-shoulder}} &= F_{\text{Earth-arm}} + F_{\text{suitcase-hand}} = W_{\text{arm}} + W_{\text{suitcase}} \\ F_{\text{hand-suitcase}} &= F_{\text{suitcase-hand}} = W_{\text{suitcase}} \end{aligned}$$

So the force your shoulder applies to your arm is larger than just the weight of the suitcase. Your shoulder has to hold up both the suitcase and your arm.

How much does this matter? To know you have to figure out how much your arm weighs. How might you estimate this? Let's say your arm is a cylinder about 10 cm in diameter and 0.8 m long. This would have a volume:

$$V_{\text{arm}} \approx (\pi * 0.05^2) * 0.8 = 6.3 \times 10^{-3} \text{ m}^3$$

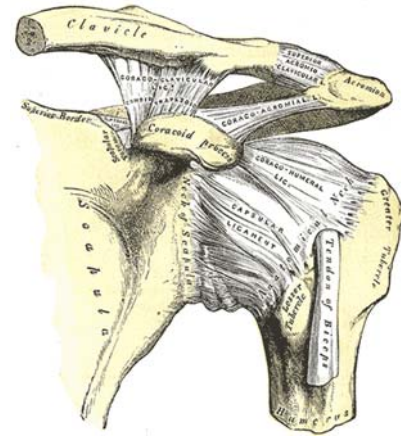
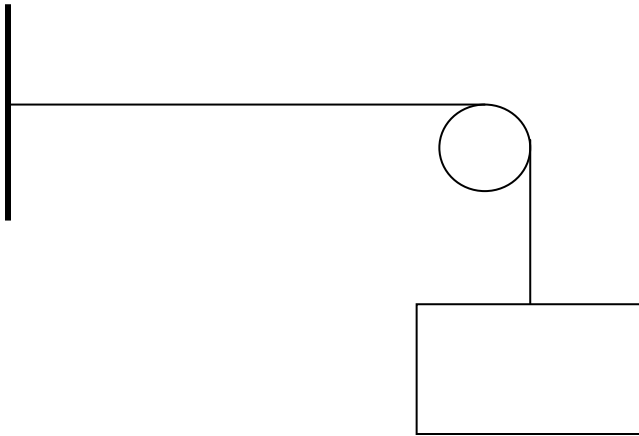
To find the mass, we multiply this volume by a density. What's the density of your arm? You are mostly made of water, so a reasonable estimate is the density of water, which is about 1000 kg/m³. So the estimate for the mass of your arm is:

$$m_{\text{arm}} \approx \text{density} * \text{volume} = 1000 \text{ kg/m}^3 * 6.3 \times 10^{-3} \text{ m}^3 = 6.3 \text{ kg}$$

The force your shoulder applies to your arm is $W_{\text{arm}} + W_{\text{suitcase}} = m_{\text{arm}}g + m_{\text{suitcase}}g = 287 \text{ N}$. To just hold up the suitcase the force would be less: $W_{\text{suitcase}} = 225 \text{ N}$.

Ropes and sending forces around corners

One important reason that ropes and tendons are important is their ability to transmit a force around a corner. Consider the following variation on the block hanging from the ceiling:



What happens here? Here the weight of the block pulls down on the rope, stretching it, putting it in tension. The nifty thing is that the rope allows this force to be transmitted around the corner, so that ultimately it is supported by the wall on the left.

This happens very often in organisms, and for our purposes this will be the most important kind of application. A good example is the human shoulder, where a complex series of tendons transmit forces generated by muscles in your back and shoulders around the flexible corner of your shoulder. Muscles pull straight along their length, while the tendons (which do not themselves generate forces) transmit the forces generated by the muscles to where they need to be applied.

Why would physicists talk about massless ropes?

Let's go back a step to the simple hanging mass problem. We found that the force exerted by the ceiling on the rope was:

$$F_{\text{Ceiling-Rope}} = W_{\text{Rope}} + W_{\text{Block}} = (m_{\text{Block}} + m_{\text{Rope}})g$$

In other words the rope isn't a *perfect* way to send a force from the ceiling to the block; it isn't a perfect force transmitter. Some extra force is needed to support the weight of the rope. How important is this flaw? The answer depends on the details of the problem. If the mass of the rope is very much less than the mass of the hanging object $m_{\text{Rope}} \ll m_{\text{Block}}$, then we can say with some accuracy:

$$F_{\text{Ceiling-Rope}} \cong W_{\text{Block}}$$

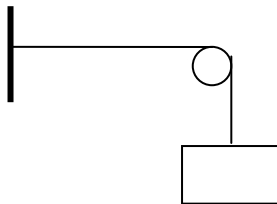
In this case the rope is a very good force transmitter. It allows the ceiling to support the block below without adding anything to the force required. What if $m_{\text{Rope}} \geq m_{\text{Block}}$? Now the force the ceiling exerts on the Rope is *at least* twice what it would be if we directly attached the mass. So in this case the rope is a really poor force transmitter.

For some of the situations we analyze we will assume that the mass of a rope is small enough to not matter. If the rope is assumed to be ‘massless’ in this way, then it becomes a perfect force transmitter. Any force applied to one end is directly transferred to the other end no matter what the circumstances. In technological cases, people always try to use ropes where this is the case, selecting a rope strong enough to support the load, but light enough to be able to transmit almost all of the force. Very often it is a reasonable approximation. The one case where this becomes very difficult is with very long ropes, like those used to sample material at the sea floor, or the cables used in the construction cranes which have become so common in Ann Arbor.

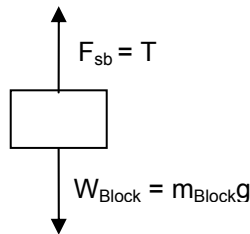
When we assume such a massless rope, the force exerted on one end is directly transmitted to the other. This force which is transmitted is what we call the ‘tension’ in the rope. So, if the tension in a rope stretched between two objects A and B is 50N, this means that object A is using the rope to pull on object B with a 50N force, and object B is using the rope to pull on object A with a 50N force. This force is perfectly transmitted by the rope, with no loss.

Tension and force transmission:

Consider the following situation. A block (m_{Block}) hangs on an essentially massless string.

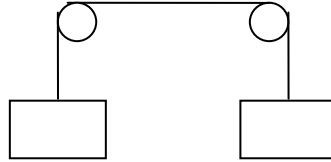


Considering the block as we did above we find:



The fact that the block is at rest implies that $T = m_{\text{Block}}g$. With what force does the string pull on the wall?

Now consider a slight variation. Instead of attaching the string to the wall, I hang a second, identical weight off a second pulley. What is the tension in the string now?



It is **still** $T = mg$. Remember, the string merely transmits a force from one end to another. It doesn't matter if the transmitted force comes from the wall or another block, it is still just transmitted.

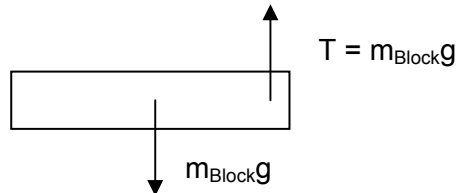
3.5 Torque and rotational statics

We have to extend our discussion of objects at equilibrium, things which aren't accelerating, to consider another kind of motion: rotation. We know from Newton's second law for translational motion that

$$\Sigma \vec{F} = \frac{d\vec{p}}{dt}$$

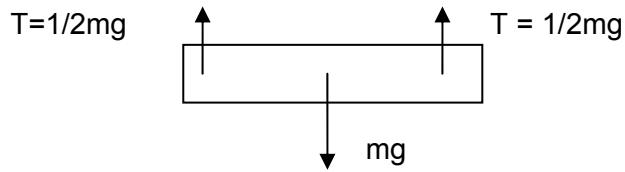
If we see that the motion of an object is not changing, so that $d\vec{p}/dt = 0$, then $\Sigma \vec{F} = 0$. This is the **first condition for equilibrium**. When the motion of an object is not changing, the vector sum of the forces on the object must be zero. For point objects this is all we need to know.

Now it is time to go beyond this, and begin thinking about how *extended* bodies will behave under the influence of forces. What happens if I have the following arrangement; a bar with a rope holding it up on one end?



This is a body for which $\Sigma F_x = \Sigma F_y = 0$. Do you think it will remain at rest? No, it will begin to rotate. Whenever a body is extended, larger than a "point" object, it is necessary to know both *what* forces act on it, and also *where* the forces are applied.

You all know from experience what we would have to do to prevent rotation in the system described above; just hang it from **two** ropes:



Why is this case stable when the other wasn't? Because here we have one force which tends to make the object rotate clockwise, and one which tends equally to make the object rotate counterclockwise. This is the basic idea of our second condition of equilibrium; the forces applied to an object at equilibrium must be applied in such a way that their tendency to make the object rotate cancels out.

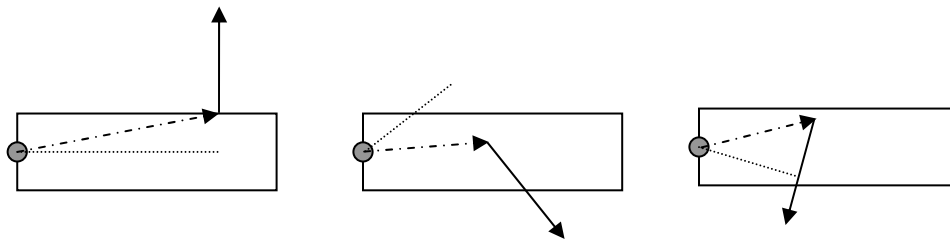
Quantifying the ability of a force to make something rotate:

It should be clear from this little discussion that the ability of a force to make an object rotate depends not only on how large the force is, but also on where it is applied. If you want to open a door (to make it rotate around its hinge) you can either apply a large force close to the hinge, or a small force far from the hinge. The direction you apply the force also affects the result. If you push on a door along a line which passes through the hinge, the door will never begin to rotate.

The number which quantifies this "ability to cause an object to begin to rotate" is called the "torque". A first definition of the torque which will generate rotation around a center C is:

$$\tau_c = r_{\perp} F$$

where r_{\perp} is called the "moment-arm" of the applied force and F is the magnitude of the force. This "moment-arm" is illustrated in the following figures:



There are three lines drawn in each picture. In each the solid line is a vector representing the force \vec{F} . It begins at the point where the force is actually applied. The dot-dashed line is a vector drawn from the center of rotation to the point at which the force is applied. We call this position vector \vec{r} . The dotted line perpendicular to the force \vec{F} in each case represents the "moment-arm" associated with this force, r_{\perp} .

One thing to note is that I can move the force forward or back along its direction and produce exactly the same r_{\perp} . A second thing to note is that the rotation produced by a force has a particular direction; it causes rotation in one direction or another. We record direction of rotation using a "right hand rule". One

way of stating this rule says that if you curl the fingers of your right hand in the direction of rotation, your thumb points in what we define as the direction of rotation.

Experience, and a second way of looking at torque

Does this definition for torque agree with experience? Let's look at a few limiting cases.

1. **\vec{F} is perpendicular to \vec{r} :** In this case, $|\vec{r}_\perp| = r_\perp$ and the torque is just $\tau = rF$. This is the familiar case of opening a door by pushing perpendicular to its surface. I can create the same torque by either pushing with a large force close to the door hinge, or pushing with a small force far from the door hinge. The torque is the product of these two.
2. **Force \vec{F} is parallel to (or anti-parallel to) \vec{r} :** In this case the force points towards or away from the center of rotation, and r_\perp is zero. The torque associated with this force is zero. Pushing straight toward or pulling directly away from a center of rotation can never cause motion about this center.
3. **Force \vec{F} is applied at the center of rotation:** This is really a subset of the previous example, and it also generates no torque and causes no rotation.

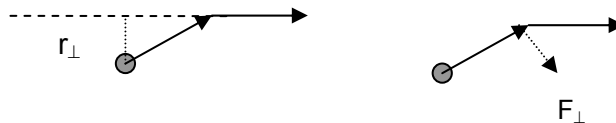
These facts suggest a second, equivalent way of looking at torque. We can take each force which acts on a body and break it up into two components, one directed along the line to the center of rotation and one perpendicular to this line. The component along the line through the center of rotation will generate no rotation. Only that component of the force which is perpendicular to the line through the center will cause rotation. So another way to determine the torque generated by a force can be written:

$$\tau_c = rF_\perp$$

So now we have two alternate ways of looking at it:

$$\tau_c = r_\perp F = rF_\perp$$

Compare these on the drawing. Both have the same force applied at the same location, and the show the two ways of visualizing the torque.

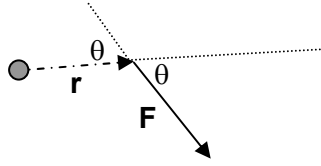


Torque and the cross-product

There is a general way to see what the magnitude of the torque will be:

$$\tau_c = r_\perp F = rF_\perp = rF \sin(\theta)$$

where θ is the angle between the vectors \mathbf{r} and \mathbf{F} , as shown in the drawing:



You can see this by noticing that $F \sin(\theta) = F_{\perp}$, or by seeing that $r \sin(\theta) = r_{\perp}$. This generic way of denoting the torque is often written in a mathematical short-hand:

$$\vec{\tau} = \vec{r} \times \vec{F}$$

This is the second of the two kinds of vector multiplication which we defined in Chapter 1. It's called a cross product, and the way we would say this equation in words is "torque is equal to r cross F". What it means is that the torque is a vector perpendicular to both \vec{r} and \vec{F} , which has a magnitude $rF \sin(\theta)$. The direction of this torque vector comes from the right hand rule we described above. Point the fingers of your right hand along r , then curl them towards F , and your thumb points in the direction of the torque. Because the cross-product of two vectors produces a vector as its result, it is also called the vector product.

Notice one essential feature of this: You cannot calculate how large a torque some force produces until you specify exactly what center of rotation you're talking. Sometimes the center of interest will be obvious, other times, like in some examples which will follow, it will not.

A Quick Summary of Some Important Relations

Weight – the pull of the Earth:

$$W = F_{\text{Earth on an object}} = mg \text{ (toward the center of the Earth)}$$

Free body diagrams:

The motion of an object is completely determined by the forces which act on it. To understand those forces, consider the object alone, completely free and separated from other objects, and worry *only* about the forces which are applied to it. Anything which does not apply a force cannot affect its motion. Drawing the object alone, and noting all the forces which act on it, is creating a ‘free body diagram’. This is the first step in solving a problem in Newtonian mechanics.

Contact forces between solids:

These forces are often divided into a component perpendicular to the interface between the two objects, the “normal force”, and a component parallel to the interface, a “frictional force”. The normal force is a passive force which takes on whatever magnitude it must to prevent one object from moving through the other.

Tension and force transmission:

Flexible materials like ropes and tendons can be used to apply forces to distant objects. In this sense they ‘transmit’ a force from one place to another, and even around corners. So long as the rope is much less massive than what it pulls, it may do this transmission very effectively.

Torque and rotational statics:

Just as force is what causes changes in motion, ‘torque’ is what causes changes in rotational motion. The size of a torque created by a force depends on both the size of the force and how it is applied. To create a large torque the force should be applied far from the center of rotation, and perpendicular to a line from that center to where the force is applied. This is quantified in the definitions of torque:

$$\vec{\tau} = \vec{r}_{\text{center to force}} \times \vec{F}$$
$$|\vec{\tau}| = r_{\perp} F = r F_{\perp} = |\vec{r}| |\vec{F}| \sin(\theta)$$

4. Understanding staying put: applying conditions for equilibrium

- 1) Two conditions for equilibrium
 - i. Weight and center of gravity
 - ii. Stability and balance
 - iii. A detailed example
- 2) Your body as a structure
 - i. Lifting a weight
 - ii. Standing on tip-toe
 - iii. Bones, tendons, and muscles
- 3) The inadequacy of two equilibrium conditions
 - i. Real problems are too complex: more information is needed
 - ii. Structures and materials
 - iii. Hooke and the forces applied by objects
- 4) Materials instead of objects; the microscopic view
 - i. Specific response of materials: stress and strain
 - ii. Keeping them straight: stress causes strain
 - iii. Relating the two in stress-strain graphs
- 5) Stiffness of materials
 - i. Young's modulus when stress-strain graphs are linear
 - ii. Stress, scaling, and the shapes of living things
 - iii. Other forms of stress: shear and pressure
- 6) Limitations of the linear stress-strain model
 - i. Strength: plastic deformation and rupture
 - ii. Non-linear materials: the stuff of life
 - iii. The J-curve and its advantages

Physics for the Life Sciences: Chapter 4

4.1 Understanding staying put: applying the conditions for equilibrium

We have seen that there are **two** conditions required for equilibrium of an extended object. The total force on the object must be zero, otherwise the object would start moving. In addition, the total torque on the object must be zero, otherwise it would start rotating.

$$\Sigma \vec{F} = 0 \quad \text{and} \quad \Sigma \vec{\tau}_{\text{every center}} = 0$$

Now there is a key feature to this second condition. It says that the sum of the torques around **every possible** center of rotation must be equal to zero. If this were not so, if there were *some* center of rotation around which the torque was *not* zero, the object would begin to rotate around that center. This is a powerful fact. When you are analyzing some situation and finding the forces and torques on a body in equilibrium, you can sum the torques around any convenient center of rotation you like and you know the sum must always equal zero. Judicious choice of which center to use can very much simplify the calculations required in many statics problems.

In this chapter, we will learn how to apply these two conditions to the study of structures like elbows, elephants, and bridges. We will see that they have much to tell us about why organisms have the shapes and sizes which they do. To begin, we need to examine just where it is that forces act.

Many of the forces we've been talking about are contact forces; a push from a finger, the pull of a rope. Each is applied at a particular place on the surface of an object. It is clear where they act, and hence easy to calculate what kinds of torques they create. But there are exceptions. One comes from non-contact forces like gravity. They don't occur only at points of contact, so where do they act? The others are distributed forces like the normal force. In a simple case, like our book on the table, it's not hard to guess how the force will be distributed. The total normal force is spread equally across the area of contact. But what if the book lies on a slope? How is the normal force distributed then?

Weight and the center of gravity

The only non-contact force we are concerned with right now is gravity. Where does gravity act? How do we determine what kinds of torques the weight of an object might produce? Gravity is an interaction through which each bit of mass pulls on every other bit of mass. So when the Earth pulls downward on our book, it applies a small force to each small part of the book. The force on each little part is determined by the mass of each little piece: $\Delta W = \Delta mg$.

Keeping track of all these little gravitational forces would be a challenging task. But fortunately there is again a powerful simplification we can use. We can always treat the force of gravity on an object as if the entire force were being applied at one particular spot called the "center of gravity". For objects near the surface of the Earth this is in the same mass-weighted average position of the object which we also call the center of mass. To find it, you multiply the mass of each little piece of the object times a vector which represents its position, add all these up, then divide by the total mass:

$$\vec{R}_{CG} = \frac{\sum_i m_i \vec{r}_i}{\sum_i m_i} = \frac{\int \rho(\vec{r}) \vec{r} dV}{\int \rho(\vec{r}) dV}$$

In this definition the sum is taken over all the little pieces which make up the object, each of which has a mass m_i . In the integral form we replace m_i with the density at the point \vec{r} multiplied by a little volume element at that point: $\rho(\vec{r}) dV$. If the object is *not* near the surface of the Earth the center of gravity may be different from this center of mass. This happens only for *really* large objects, like the Moon. In our considerations of living things we needn't worry about this.

For homogenous (with the same density everywhere) and symmetrical objects, the center of gravity is always in a obvious location. For a sphere it is at the center, for a hoop at the center, for a square at the center, a cube at the center, etc. If the object is either not homogeneous or not symmetrical the center of gravity will be "pulled" toward the parts which are more massive.

For example, consider the CG of a 30 cm long bar with a spherical 10 kg mass on one end and a spherical 20 kg mass on the other end. We'll ignore the mass of the bar. Where is the center of gravity for this object? Taking the origin to be the location of the 10 kg mass, we have:

$$\vec{R}_{CG} = \frac{\sum m_i \vec{r}_i}{\sum m_i} = \frac{10\text{kg} \times 0\hat{x} + 20\text{kg} \times 0.3\hat{x}}{10\text{kg} + 20\text{kg}} = 0.2\hat{x}$$

Rather than being in the center of the barbell, the center of gravity is pulled toward the larger mass, and is instead two thirds of the way down from the end. If you wanted to balance this barbell on one finger, you should place the finger at just this point, and of course your finger had better be rather strong.

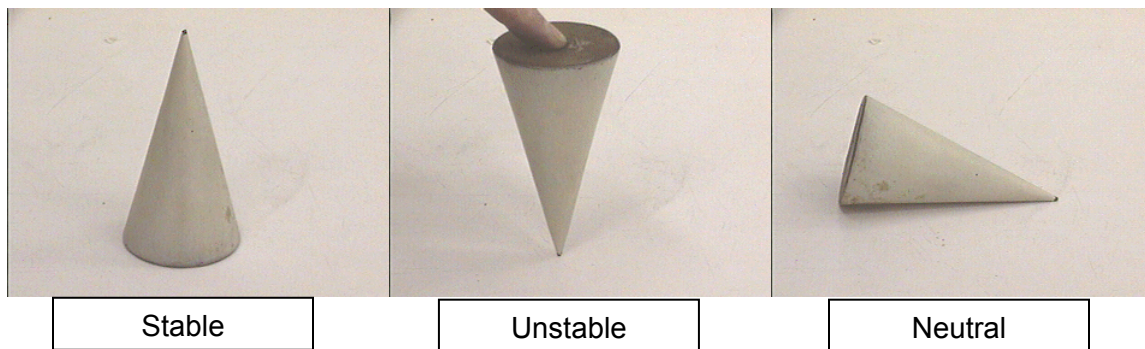
There is an important lesson here: Because gravity acts equally on all parts of an object, it cannot cause rotation around the center of mass. It can, of course, cause rotation around other centers. Contact forces, on the other hand, always act on particular parts of the body, and hence can, in principal at least, generate rotation around any point on an object.

An application: stability and balance

A fine application of the idea of equilibrium is stability and balance. Thinking about this is especially useful as it helps to illustrate the different kinds of equilibrium. There are three kinds of equilibrium:

1. Stable equilibrium: if the object is slightly disturbed it will return to equilibrium
2. Unstable equilibrium: if the object is even slightly disturbed it will move far away from equilibrium
3. Neutral equilibrium: a small displacement leaves the object in a new equilibrium position

These three are illustrated by thinking about a cone. When the cone is standing on its bottom, it is in stable equilibrium. Tip it a bit and it falls back into place. If it's standing on its tip, it is in unstable equilibrium. Tip it a bit and it falls over completely. If you lay the cone down on its side it is in neutral equilibrium. Roll it over a bit and it just lies there.



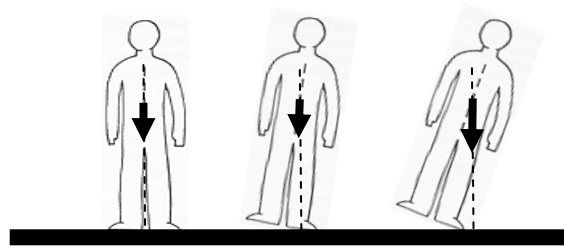
For simple cases of balance, like those which apply to many organisms the following rules apply:

1. If a tilt away from equilibrium *raises* the center of gravity, the object is in stable equilibrium
2. If a tilt away from equilibrium *lowers* the center of gravity, the object is in unstable equilibrium
3. If a tilt away from equilibrium leaves the height of the center of gravity unchanged, the object is in neutral equilibrium

This distinction arises because of the torque exerted on the object by the force of gravity. If you're raising the center of gravity when you tip it, the force of gravity will tend to pull it back down. If you're lowering the center of gravity, the force of gravity will tend to pull it away from where it started.

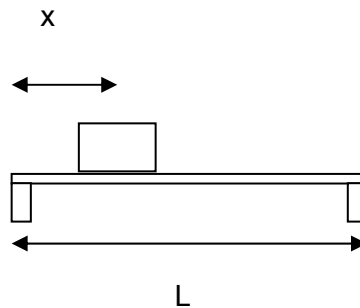
When will an object "tip" over? Any time the center of gravity of an object is not above the supporting surface of the object it will tip. Once the center of gravity moves past the support point, gravity exerts a torque which tends to tip the object over.

Consider these pictures to get an idea of how this works. As we tip this stiff little person over he is, at first, stable. We're still raising the center of gravity. Eventually, the center of gravity moves outside the support point, and is now moving downward. This is unstable.

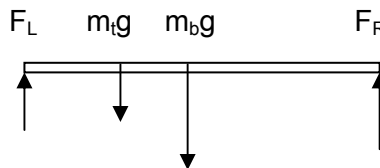


A quantitative example of rotational equilibrium: a truck on a bridge

A first quantitative example comes from analyzing a truck driving over a bridge supported on its two ends. What might we like to know about this? Imagine we know the weight of the truck, and we want to know how much force must be applied by the right and left hand supports as the truck moves across the bridge. This is just the sort of thing an engineer or an architect might need to know to make sure the bridge is safely constructed.



We need to draw a free body diagram for something. Let's begin by considering the forces on the bridge slab; the road itself. There is an upward force on each end, a downward force of the weight of the bridge, and a downward force due to the weight of the truck.



We can find the magnitudes of F_L and F_R by using the equations of equilibrium. There are no forces in the horizontal direction, so we sum the forces in the vertical direction, taking up to be positive.

$$\Sigma F_y = F_L - m_t g - m_b g + F_R$$

Let's take the left end of the bridge and sum the torques around this point. We need to choose a sign convention for rotation as well: we will take the counter-clockwise direction to be positive. If we do that, we get:

$$\Sigma \tau_{\text{Left end}} = -m_t g x - m_b g \left(\frac{L}{2} \right) + F_R L$$

Why do I choose the left end of the bridge as the center of rotation? After all, I could calculate the torques around *any* center and they must always be equal to zero. This choice is made purely for convenience. By picking a center through which one of the two unknown forces passes (F_L goes through this spot) I know that only one of the unknown forces (F_L and F_R) will appear in the torque equation I obtain. Making this choice just makes the algebra a little simpler than it would otherwise be.

Solving the torque equation yields:

$$F_R = m_t g \frac{x}{L} + \frac{1}{2} m_b g$$

And plugging this back into the first equation gives us the other unknown force:

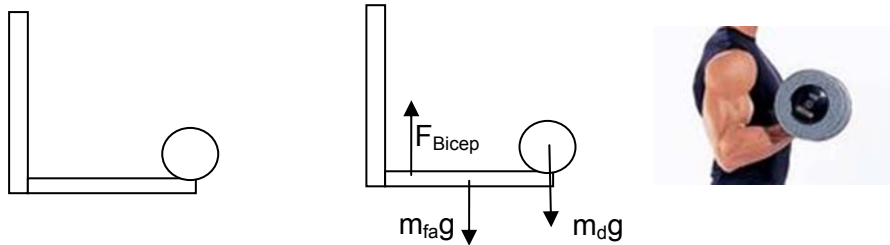
$$F_L = m_t g \left(1 - \frac{x}{L} \right) + \frac{m_b g}{2}$$

What do these equations tell us? The upward force exerted by the right hand bridge support is half the weight of the bridge plus a fraction of the weight of the truck that varies as it drives across the bridge. When it is over the left hand support, all of the truck's weight is supported on the left, when it reaches the middle, half of it is supported by each, and when $x=L$, all the weight is supported on the right. The upward force exerted by the left hand support makes up for the rest of the weight of the bridge and truck. Added together, $F_L + F_R = m_t g + m_b g$, all the time. This is a nice answer, which we might have anticipated, worked out using our equilibrium conditions.

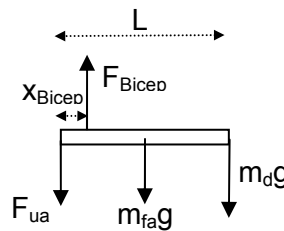
4.2 Your body as a static structure: weight lifting

Your bodies, and the structures of other living things, obey these same principles of statics. These simple ideas, that in a static situation the sum of forces and sum of torques must be zero, are all it takes to understand a lot about how your body works. Let's start with a simple example: holding a dumbbell with your forearm held horizontal.

Imagine that the mass of your forearm is m_{fa} , the length of your forearm is L , and the mass of the dumbbell is m_d . How do you hold this weight up? You have a bicep which attaches to your forearm just a few centimeters from the joint in your elbow. If you poke around the inside of your elbow now, you can feel the tendons which connect this muscle to your forearm. We will call this distance x_{bicep} . So the picture looks something like the diagram in the center.



Remember, the only way to analyze forces is to consider **a single object in a free body diagram**. So let's look just at the bone in the forearm:



If you just had the three forces we mentioned (weight of forearm, weight of barbell, and upward bicep force F_{Bicep}) your forearm would rotate clockwise. So, there must be another force. What is it? It's the force of the end of your upper arm bone pushing **down** on the end of your forearm: F_{ua} . This push-pull arrangement is always present when your body supports weight. Your body works by pulling with muscles and pushing with bones. The combination of the two is what allows your full range of movement, and you need them both: muscles must have something to pull against.

To determine the size of these various forces, we have two facts to work with:

$$\Sigma \vec{F} = 0 \quad \text{and} \quad \Sigma \vec{\tau} = 0$$

Here there are only y forces, so:

$$\Sigma F_y = F_{Bicep} - F_{ua} - m_{fa}g - m_dg = 0 \quad \text{or} \quad F_{Bicep} = F_{ua} + m_{fa}g + m_dg$$

If we sum the torques around the end, we find:

$$\Sigma \tau = xF_{Bicep} - \left(\frac{L}{2}\right)m_{fa}g - Lm_dg = 0 \quad \text{or} \quad F_{Bicep} = \left(\frac{L}{x}\right)\left(\frac{1}{2}m_{fa}g + m_dg\right)$$

What does this mean for people of different sizes? If I make a person larger, both L and x are likely to increase in a similar way. So how does the force they must apply change? It wouldn't change at all. This

would suggest that lifting a dumbbell of mass m_d would require the same force for a small person or a large person.

Be careful though, for we have buried in this statement an assumption of isometry. Isometry means that during a change in size, all dimensions of the person change by exactly the same factor. This is another way of saying the size change is isomorphic, and preserves the shape. The *shape* of the person is the same, only the *size* of the person is different. If this is violated, let's say because one person is short and stocky and the other is tall and lanky, the simplest scaling will not hold.

For the short and stocky person, L/x will be relatively small. For the tall and lanky person, L/x will be relatively large. Hence it will be easier (require less bicep force) for the small and stocky person to hold up the barbell than the tall and spindly person. This is in perfect accord with our sense that short stocky people seem stronger than tall and lanky people. Short, stocky people actually can lift more with muscles of the same intrinsic strength. Understanding the physics of statics allows us to appreciate these simple scalings, and to explain a lot of what we know about how people move.

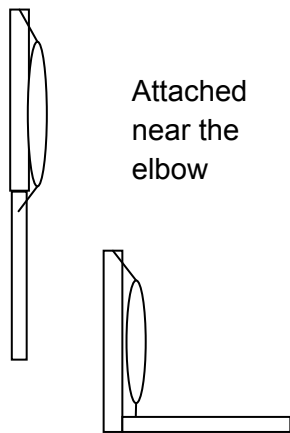
Returning to the first equation:

$$F_{ua} = \left(\frac{L}{x}\right) \left[\frac{1}{2} m_{fa} g + m_b g \right] - m_{fa} g - m_d g = \left(\frac{L}{2x} - 1\right) m_{fa} g + \left(\frac{L}{x} - 1\right) m_d g$$

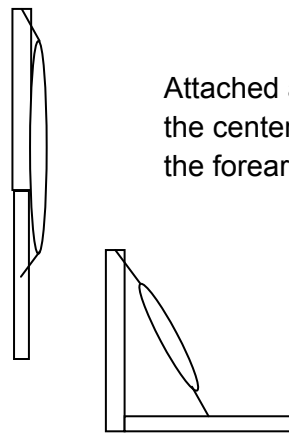
Note that both F_{Bicep} and F_{ua} go to ∞ when x goes to zero. If your body is going to function, the tendon from your bicep must be attached some small distance from your elbow. The farther away it is attached (the larger x is), the easier it is to lift something like this dumbbell. Both F_{Bicep} and F_{ua} become smaller. The closer to the joint your bicep is attached, the harder it is to lift a load. Given this, you might ask why it is that your bicep is not attached farther out along your forearm.

As is usual in systems arrived at through natural selection, this design represents a compromise. Moving the connection of the bicep to the forearm farther out would make it possible for you to lift larger loads with the same muscle. So why hasn't this happened? What's the drawback?

A bicep attached farther out, say at the center of the forearm, would have to contract *much* more to take your arm from stretched out straight to fully bent. The ability of muscles to contract along their length is limited by the biology of the muscle cells, and in the end this limitation more than makes up for the increased force required when the muscle is attached closer to the elbow.



Attached near the elbow

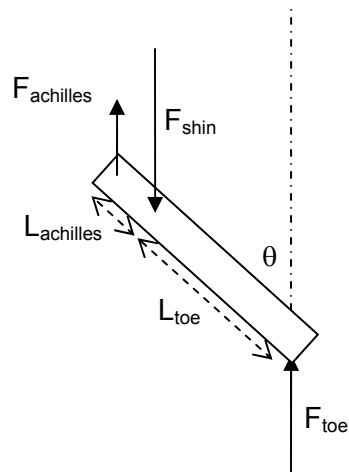
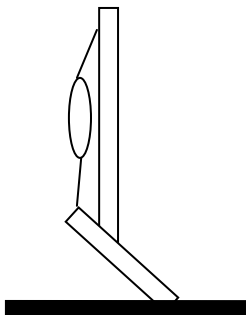


Attached at the center of the forearm



Standing on tiptoe

Here's another very common situation we can understand with statics principles: standing on tiptoe. How do the forces work for this?



We'll start with a few simplifying assumptions. First, ignore the weight of the foot itself. This is probably OK because your body weighs so much more than your foot. Second, assume all three forces shown act straight up and down. In reality the forces applied by your shin (F_{shin}) and your Achilles tendon (F_{Ach}) don't act quite straight up and down, but they nearly do.

Let's sum the forces:

$$\begin{aligned}\Sigma F_x &= 0 \\ \Sigma F_y &= F_{\text{achilles}} + F_{\text{toe}} - F_{\text{shin}} = 0\end{aligned}$$

And sum the torques around the point where the shin bone contacts the foot:

$$\Sigma \tau_{\text{around shin}} = F_{\text{toe}} L_{\text{toe}} \sin(\theta) - F_{\text{achilles}} L_{\text{achilles}} \sin(\theta) = 0$$

These two (useful) equations contain three unknowns. So we can't solve them with more information. What information is there? We know that the floor must support your total weight. If we assume you're standing on tiptoe with *both* feet, then each must support half your weight, and we know that:

$$F_{\text{toe}} = \frac{m_{\text{you}} g}{2}$$

Using this additional fact we can find:

$$F_{\text{achilles}} = \frac{L_{\text{toe}}}{L_{\text{achilles}}} F_{\text{toe}} = \frac{L_{\text{toe}}}{L_{\text{achilles}}} \frac{m_{\text{you}} g}{2}$$

There are some interesting things to note. First, the force applied by the Achilles tendon is independent of the angle of your foot. This is a little surprising, but if you try it out you'll see it's approximately true. It is nearly as easy to stand on tiptoe no matter what angle your foot is at, and you can move up and down with very little additional effort.

Second, the force applied by the Achilles tendon is larger than half your weight, usually by quite a substantial amount. Just holding a ruler up to my foot I find $L_{\text{ach}} \cong 5 \text{ cm}$ and $L_{\text{toe}} \cong 20 \text{ cm}$. So the upward force applied by each of my Achilles tendons is about four times as large as half my weight, or twice my total weight!

This is why your calf muscles are probably substantially larger than your biceps. You rarely lift twice your weight with your biceps, but you do it all the time with your calves, with every step you take.

Bones, tendons, and muscles

Large land animals (mammals, reptiles, and birds) are constructed of three primary mechanical components; bone, tendon, and muscle. Bones provide the primary support mechanisms which enable you to stand up against the continual downward pull of gravity. You might fruitfully picture yourself as a skeleton of bone with a lot of flesh and organs hanging from it.

Like most biological tissues, bone is both complex and various. In general, 50% or more of bone mass is made of calcium phosphate, a hard mineral which gives bone its strength. Bone is very strong in compression; which is how it is typically used in your skeleton. Muscles usually pull mostly along the long axis of the bones to which they attach.

This skeleton is held together, pulled one way and the other, by an elaborate set of muscles attached to the bones by tendons. Muscles are the only active agents; on receiving the right stimulus, these cells can contract along their length, generating a tensile force. The tendons then transmit these forces to the bones of your skeleton.



The rest of your tissues, all the organs like the brain, liver, eyes and so on, are, mechanically at least, just along for the ride.

4.3 The problem with our two equilibrium conditions

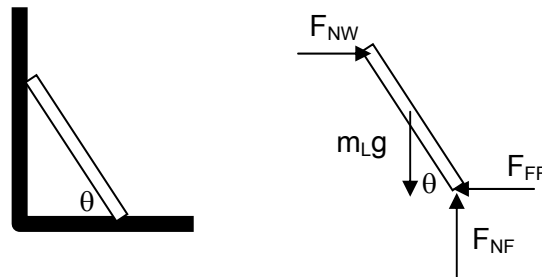
In some cases, the two rules for equilibrium allow us to fully understand how structures and organisms stand up to the pull of gravity. But as we will see, these two rules are almost always not enough to completely figure out what's happening. To fully understand organisms and other structures, we need to understand how the materials they are made of apply forces. How can a bone, or a table, or a string, apply a force? Answer that and you will have laid out the crucial piece of information needed to analyze equilibrium.

What's the problem with our equilibrium conditions? The condition for equilibrium requires:

$$\Sigma \vec{F} = 0 \quad \text{and} \quad \Sigma \vec{\tau}_{\text{any axis}} = 0$$

We have noted these conditions before when we first started to consider rotational motion. Since each of the two is a vector equation, there are really six independent equations here. Three state that the sum of the forces in each of the three directions is zero, and three state that the sum of the torques around each of three perpendicular axes is zero.

Let's try a new example to remind you how this works. Consider a ladder resting on a floor which exerts a frictional force on it. The ladder leans on a completely slippery frictionless wall. Where the ladder rests against the floor, there is both a normal force perpendicular to the floor and a frictional force along the floor. Where the ladder touches the wall, there is *only* a normal force.



How large is the frictional force applied by the floor on the ladder? Use the constraints:

$$\Sigma F_{\text{horizontal}} = F_{NW} - F_{FF} = 0 \quad \text{and} \quad \Sigma F_{\text{vertical}} = F_{NF} - m_L g = 0$$

There are three unknowns here, and two equations. So while these nicely let us find the normal force exerted by the floor, they leave us with no knowledge of the normal force exerted by the wall or the frictional force exerted by the floor.

$$F_{NF} = m_L g$$

What about the torques? Let's calculate the sum of the torques around the point at the top. Defining counterclockwise rotation as positive, I have:

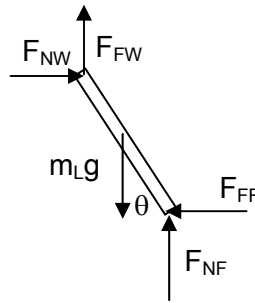
$$\Sigma \tau_{\text{top}} = F_{NF} L \cos(\theta) - m_L g \left(\frac{L}{2} \right) \cos(\theta) - F_{FF} L \sin(\theta) = 0$$

Substituting in the value for F_{NF} obtained above and dividing by L , we get:

$$m_L g \left(\frac{1}{2} \right) \cos(\theta) - F_{FF} \sin(\theta) = 0 \quad \text{or} \quad F_{FF} = \frac{m_L g}{2 \tan(\theta)}$$

As the angle gets smaller, you're going to need a larger and larger amount of friction to prevent the ladder from slipping. Your intuition will tell you that there's a limit here. If you try to stand the ladder up so that it is too nearly horizontal it will slip. No big surprise in this.

Now let's consider a just *slightly* more complicated case. The same situation, but now imagine that there *is* some friction with the vertical wall. Of course there would be in any real case.



How are the equations changed?

$$\Sigma F_x = F_{NW} - F_{FF} = 0 \quad \text{and} \quad \Sigma F_y = F_{NF} + F_{FW} - m_L g = 0$$

And if we still sum the torques around the top, we still have

$$\Sigma \tau_{top} = F_{NF} L \cos(\theta) - m_L g \left(\frac{L}{2} \right) \cos(\theta) - F_{FF} L \sin(\theta) = 0$$

But now I can no longer identify F_{NF} with $m_L g$, because part of the weight may be supported by friction with the wall! What's going on here? We now have three equations to work with, and four unknowns:

$$F_{NF}, F_{FF}, F_{NW}, \text{ and } F_{FW}$$

We know mathematically that this system of equations *cannot* be uniquely solved without more information. You might think of it as having infinitely many solutions. Imagine that you pick a value for F_{FW} (or indeed any one of the four forces). Given this value, all the other forces are determined from the equations above. But without knowing more, *all* of these infinitely many solutions are allowed by the laws of physics invoked above, and nothing we have done will tell us which of the many possibilities will actually occur.

We can illustrate this ambiguity by considering two extremes. If I take the ladder and cram it down into the corner, there will be a large frictional force at the top of the ladder pushing it up (as in the picture above). This will reduce the frictional force pushing it to the left on the bottom. If, on the other hand, I pull up some on the ladder, trying to lift it away from the corner, there will be a downward frictional force at the top preventing it from sliding up. In this case I will have to increase the inward frictional force at the bottom. In other words, there are many possible answers to this problem, and without knowing more than just the two equilibrium conditions we can't tell exactly what the forces will be.

This is quite generally the case in statics problems. At best, you have only the six equations which the equilibrium conditions provide, and in most cases this will not be enough to determine the forces in an object. Another simple example is a cow. A cow stands on four legs, each of which supplies an upward

force. In this case you can use one force equation (up and down) and two torque equations (tilting side to side and front to back) to constrain things, but there are still four forces, so you're out of luck.

A more extreme example is the suspension bridge. Here there are three equations, and potentially thousands of forces you need to know. If you're a bridge designer and you incorrectly determine even one of these forces, the entire bridge may fail.



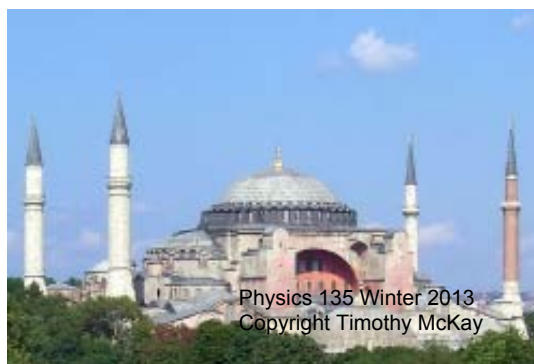
How can such a problem be solved? Nature solves such problems without difficulty. When you assemble a bridge like this it has particular forces in each element. When a cow stands in a field, there are in fact four particular forces on its legs. Something happens to determine which combination of forces (among the many which the conditions for equilibrium allow) will actually occur. What happens, and what additional information is needed to *predict* what the many forces will be?

The answer lies hidden in the ways objects actually supply forces. Remember, an object like the leg of a chair or the bone in your leg can only apply a force when it distorts a little, when it changes shape. To understand a structure like your chair, you have to know exactly *how* the chair changes shape when you apply a load to it. It's no longer enough to assume that its shape is fixed, since it actually never is.

Structures and Materials

Understanding how objects support loads is an ancient endeavor. People have always wanted to be able to build things which will stand up, rather than collapsing on their heads. Over the centuries, a tremendous amount of empirical knowledge about how structures behave was built up. This empirical knowledge was what enabled the construction of such magnificent buildings as the Hagia Sophia in Istanbul, the Taj Mahal in India, and the Cathedral at Chartres in France. Amazingly, a theoretical understanding of how structures support loads was developed only *very* much later. Some important aspects of how even common materials support loads were not understood until well into the 20th century.

The understanding of structures developed to enable people to construct buildings and machines has, more recently, been put to use in understanding living structures. The same principles which govern how buildings and bridges stand up apply as well to giraffes and antelopes. There is one qualitative difference we will emphasize later: man-made structures tend to be made of stiff, inflexible material, while nature often uses stretchier, more malleable stuff. As we'll see, there are a lot of reasons evolution settled on this solution. Indeed more and more manmade objects (like plastic bumpers and airbags) share this sort of



Physics 135 Winter 2013
Copyright Timothy McKay



flexibility.

A Little History: Hooke and how objects support loads

Newton understood that when you hang a mass from a string its weight must be supported by a tension in the string. Newton knew *what* happened in a case like this, but didn't discuss *how* the inanimate string could apply a force to the block. Perhaps he wasn't interested enough in such practical things. Robert Hooke, Newton's contemporary (1635-1703) and sometime competitor, was intensely interested in practical things. He built the first air pump (for Robert Boyle), discovered the diffraction of light and used it to promote a wave theory of light, and was the first person to explicitly note the expansion of materials when they are heated.

Hooke also published a book called "Micrographia", an enormously influential (and beautiful) series of images obtained through his microscope. In these images he revealed, among many other things, that a drop of water is alive with microorganisms, that the eyes of a fly are compounded of thousands of individual ommatidia, and that living tissues are made of many tiny "cells". In fact Hooke invented the name. In many ways, this book was a starting point for the life sciences. You can look at some of this exceptionally beautiful book online at:

<http://archive.nlm.nih.gov/proj/ttp/flash/hooke/hooke.html>

For the topic at hand, Hooke's most important discoveries had to do with how objects support loads. His main realizations are part of what is now called Hooke's law. He found that objects can support loads only by yielding to them. If you push on an object, it will be distorted; squashed by some amount. He noted that most solids are elastic as opposed to plastic. This means that if you squash them with modest forces and let them go, they spring completely back to their original shapes. Elastic objects (like rubber) do this. Plastic objects (like clay) change shape permanently when you distort them.

Finally, Hooke constructed a first quantitative model for the response of solids to loads by noting that the amount of deformation is, in many solids, proportional to the load you place on them. These observations are encoded in a general form as a very simple form called "Hooke's Law":

$$F_{\text{applied to the object}} = k\Delta x$$

When you apply a force to a solid, it is squashed by some amount Δx in the direction of the force which you apply to it. Newton's third law tells us that when you push on this solid, it will push back on you with an equal and opposite force which resists the deformation. The solid is trying to return to its original shape.

$$F_{\text{applied to the object}} = k\Delta x \qquad F_{\text{object pushing back}} = -k\Delta x$$

When you think about Hooke's law, make sure you keep track of which element of this third law pair you're interested in. One is the force applied *to* the object, the other is the force applied *by* the object.

In Hooke's law, the constant " k " tells you how hard large these forces will be for a given deformation Δx . Each object has a different constant, which you can determine by applying a force F and seeing how far the object yields (Δx). If the constant k is small, the object is easy to deform; you would call it flexible. If the constant k is large, the object is hard to deform; you would call it stiff.

The signs in these equations remind you that when you stretch an object outward, so that Δx is positive, the force you exert is in the positive direction, while the force the object exerts on you will be negative, in the opposite direction. If you squash the object inward, so that Δx is negative, the force you exert on the object will be negative, while the force the object exerts on you will be positive.

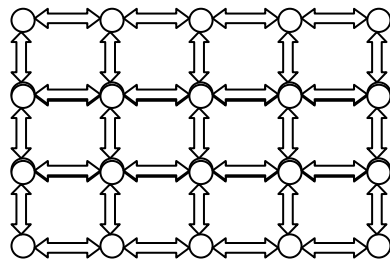
4.4 Limitations to Hooke's picture: it only works for individual objects

The problem with Hooke's law for practical purposes is that it makes no predictions about what k will be for a particular object. If I'm interested in one particular object I can measure its spring constant, and then predict exactly how it will deform under a load. But this is not too practical in making a new building. It suggests that you must build it first before you can see whether it will work. This was, in fact, how cathedrals were built in Europe. A fair number were constructed, fell down, and then were rebuilt with more extensive supports.

The problem is that Hooke focused on individual objects, rather than the materials of which they were made. He could tell you about an individual rope or beam, but was unable to make predictions about a new one. To give a specific example, Hooke could not answer the question: "I know the constant k for this particular beam. What will the new constant k' be if I make a beam of the same material which is twice as long?"

Focusing on materials instead of objects

The key to answering this question is to focus on materials. To begin, let's think about a toy model of what a solid object is like. Solids are built of atoms more or less locked in place by bonding with their neighbors. You can think of this as an array of atoms held together by rather stiff springs.



If I apply a force which is spread out over the top of this solid, it will have to squash all of these springs at the same time. If I apply the force to just a little spot, it will have to squash only a few springs. This is easier to do, so that same force applied to a small spot will squash things more; it will create a larger distortion.

To quantify this, consider a 1 and 2 spring model, where each individual spring has the same spring constant k . This would be the case, for example, for each of the little springs that connect the atoms in a solid. If I apply a force F to one spring, it is compressed a distance $\Delta x = F/k$. If I put two such springs in parallel, the combined spring constant will be $2k$, and the same force F will compress them only half as much: $\Delta x_{2\text{springs}} = F/2k$.

You can see from this example that what matters is the force *per spring*. If I measure the force per spring, all objects which are made of this material (whether big or small) will behave the same way. Now we can't actually measure the "force per spring" unless we know exactly how far apart the atoms are. So instead, we account for this by asking whether the force is spread out over more springs or fewer. We can do this by measuring the area over which the force is applied. If we double the area, the force will be applied to twice as many springs. If we halve the area, the force will be applied to half as many springs.

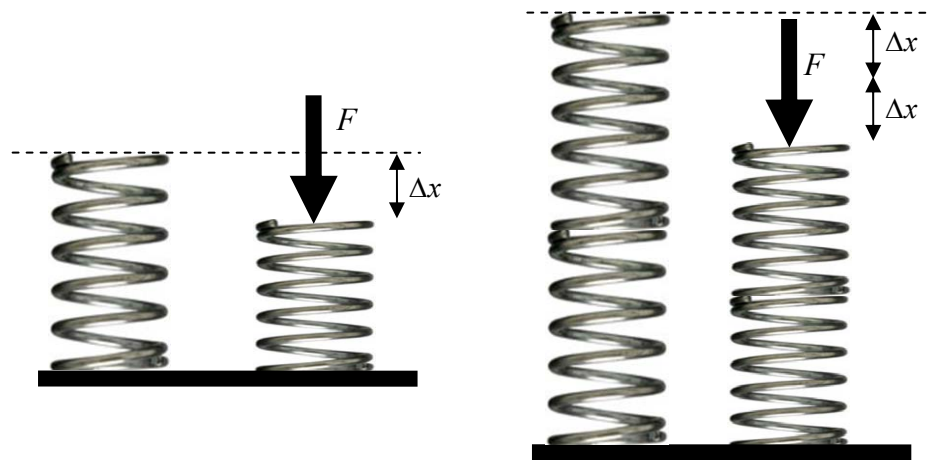
So instead of just measuring the force applied to an object, what we will care about is the force per unit area. In this application the quantity force per unit area is called the **stress**, and it's defined as:

$$\text{Stress} = \frac{\text{Force}}{\text{Area}} = \frac{F}{A} = \sigma$$

Here F is the total force, and A encodes the area (proportional to the number of atoms) over which this force is applied. The symbol σ is usually used for this quantity. Notice that stress is really the same as the more familiar pressure. There's no clear distinction between the two, though the term pressure is used more often in cases of fluids, as for air pressure or hydrostatic pressure.

Now think about the displacements in a real material. When we press on a single spring with a force F it compresses by an amount $\Delta x = F/k$. Now imagine I have two springs stacked on top of one another. We would say these springs are 'in series' with one another. In this case *each* spring will compress by a distance Δx , because the same force will be transmitted through both, one after the other. Since each spring will be squashed, the combination of two springs will compress twice as much as one. This is somewhat tricky, so think about it carefully.

What matters for displacement is the compression *per spring*. Again, we can't really do this 'per spring' without knowing exactly how far apart the atoms are, so instead we use the thickness of the material as a way of tracking whether there are more or fewer atoms. If the material is twice as thick, there will be twice as many atoms and the material will compress twice as far.



To keep track of this, we measure not just the change in length of the material (ΔL) but the *fractional* change in length, which is called the **strain**. It is defined as

$$\text{Strain} = \frac{\Delta L}{L} = \varepsilon$$

Here ΔL is the total displacement of the object, how much its length changes, and L is the object's total length. The symbol ε (the Greek letter epsilon) is usually used to denote strain.

Stress and Strain: avoiding the obvious confusion...

Using the words “stress” and “strain” brings us back to the problem of using ordinary words to describe physics concepts. In everyday language, both stress and strain mean something similar. But in physics they mean very particular, quite different things.

The stress is a measure of how much force per unit area is applied to an object. It has units of N/m^2 . The strain is our measure of how much an object is distorted by the stress which is applied to it. Strain, measured as the ratio of the distortion ΔL to the total size L , is dimensionless.

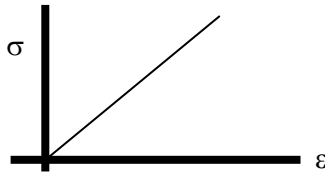
You'll have to find a way to remember, unflinching, the difference between these two. One way is to practice the mantra: “stress causes strain...stress causes strain...stress causes strain...” until you can't think of it any other way. If you invent some other good mnemonic for this please share it with your peers and your instructor. It's been a problem for students throughout the ages. Maybe you can solve it.

4.5 Stress causes strain: a material dependent version of Hooke's Law

For every different kind of material, there will be some relation between stress and strain. If you apply a stress σ , you will measure a strain ε . Very often it is the case, at least for small stresses, that the stress and strain will follow a simple, linear relation. Now, although you would probably do the experiment by applying a stress and measuring a strain, people usually graph this relation by showing the stress as a function of the strain.

This harks back to Hooke, who wrote his law $F_{\text{applied to the object}} = k\Delta x$. He wrote it as if the force was a function of the distortion Δx , rather than the distortion a function of the force. Echoing Hooke, we write:

$$\text{stress} = \text{constant} \times \text{strain} \quad \text{or} \quad \sigma = E \varepsilon$$



Note that this equation does not describe an object; it describes a material. If you make an object of a material you have studied in this way, you can predict how it will behave. Remember too that this linear relation definitely does not describe the relation between stress and strain in *all* materials. It is true that many materials have such a linear relation, at least for small stresses. But quite a few, and especially materials which are used by living organisms, do not. We'll talk more about this later.

The ideas of stress and strain, very important ones, were first discussed by Thomas Young, a remarkable 19th century scientist who we will encounter later introducing the concepts of work and energy to physics. The focus on stress and strain, rather than force and distortion, for the purpose of understanding structures was later codified and developed by a series of French theoreticians like Augustin-Louis Cauchy.

What is the constant 'E' in this equation?

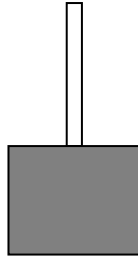
$$\text{constant} = \frac{\text{stress}}{\text{strain}} = E = \frac{\sigma}{\varepsilon} = \frac{F/A}{\Delta L/L}$$

This constant is called the 'modulus' of a material. Modulus just means "little measure" and is called this because it is a property of the material. Since the strain is unitless, the modulus has the same units as the stress, N/m². For the simple case where the stress is either tension or compression (stretching or squashing) the modulus is called Young's modulus, and is usually denoted with a capital E (though sometimes a capital Y is used).

Some Young's moduli:

Material	E
Tendon	0.6x10 ⁹ N/m ²
Oak	14x10 ⁹ N/m ²
Marble	50x10 ⁹ N/m ²
Steel	200x10 ⁹ N/m ²
Diamond	1200x10 ⁹ N/m ²

Let's work through an example to see what Young's modulus means in practice. Consider the simple case of a rod in tension. We can create a situation like this by hanging a large weight, with mass M from a rod with length L and cross-sectional area A . We will ignore the weight of the rod.



Stress in rod = modulus \times strain

$$\frac{Mg}{A} = E \frac{\Delta L}{L}$$

$$\Delta L = \frac{Mg}{A} \frac{L}{E}$$

If you double L , you double ΔL . If you double A , you halve ΔL .

Consider wood, which has a Young's modulus $E \approx 10^{10} \text{ N/m}^2$. If I hang a 1000 N weight on a 1 cm x 1 cm x 10 cm piece, I will get a stretching:

$$\Delta L = \left(\frac{F}{AE} \right) L = \left(\frac{1000 \text{ N}}{(0.01 \text{ m})^2 (10^{10} \text{ N/m}^2)} \right) 0.1 \text{ m} = 1 \times 10^{-4} \text{ m} = 0.1 \text{ mm}$$

A 1000 N weight is something with a mass of about 100 kg, a good bit more massive than a typical college student. So even if you hang from a piece of wood only 1 cm x 1 cm square and 10 cm long, you will stretch it by less than 0.1 mm.

These very large Young's moduli imply that it is quite difficult to either stretch or compress a solid. It's this property that human engineers like, and why they choose such stiff materials for constructing things. An engineer likes to know that a certain part will always stay the same size (at least very nearly) regardless of what you do to it. As we will see in a bit, most biological materials are quite different from this, and have many attractive features that our engineering materials often lack.

Stress, strain, and the scaling of living things

Now that we have recognized the importance of stress for understanding structures, we can resolve an enormous range of interesting questions about life. One of the most obvious differences between large organisms (think dogs to elephants) and small ones (think aphids to birds) is sturdiness. A rhino is a solid thing, with thick legs held straight beneath its body. If I showed you a picture of such a beast, even without any scale information, you'd know this was a massive creature. A spider, by contrast, has absurdly long spindly legs, stretching way out to the sides. You never see such ungainly structures in creatures as large as rhinos.

The reason for this difference lies in scaling laws. Imagine a simple rhino model with four straight legs. As we have often argued, the volume (and therefore mass and weight) of the rhino will vary like the size of the rhino cubed. Meanwhile the area of its legs will vary like the size of the rhino squared. This implies that the stress in the rhino's legs, which is force per unit area, will vary like:

$$\sigma = \frac{F}{A} \propto \frac{\text{size}^3}{\text{size}^2} = \text{size}$$

If you increase the size of a rhino by ten times, the stress in its leg bones will increase by a factor of 10. Remembering that the material making up the bones stays the same, you can see how just making an organism larger becomes risky very fast.

For this reason, large animals have different shapes from small ones. They have evolved proportionately thicker legs. They have also adopted postures which tend to keep the legs straight and directly under the body. On the other end of the size spectrum, tiny organisms have no trouble at all keeping the stress in their limbs low. As a result, they can adopt a much wider array of shapes, with long spindly legs that allow them to walk uninterrupted over the extremely varied terrain they see at their size.



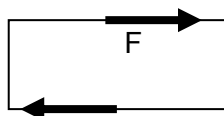
Can you tell which is smaller?



You can see that an appreciation for how materials support loads, of the importance of stress, helps us to understand a lot about the diversity of form we see in living things. Organisms don't simply choose the shapes they take. These forms are, in a very real way, imposed on them by physical constraints. There is *much* more we could say on this topic¹.

Other kinds of stress and strain

There are several other kinds of stress and strain. Since they affect the "springs between the atoms" differently, they have to be accounted for differently. The first new kind of stress is called "shear". Shear is what happens when you try to shove the top of something sideways relative to the bottom. As an example, consider laying a textbook on the table, then sliding the front cover of a textbook to the left while pushing the bottom to the right.



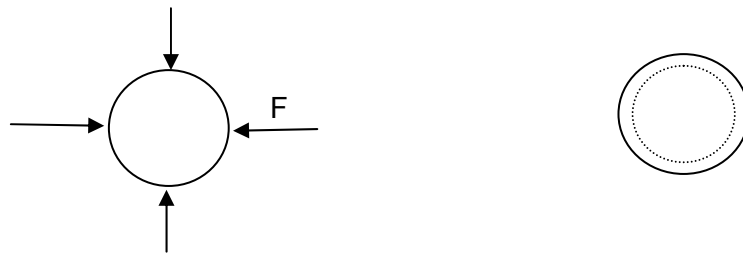
$$\text{Stress} = \frac{\text{Force}}{\text{Area of top}} \qquad \text{Strain} = \frac{\Delta x}{L}$$

These are called ‘shear stress’ and ‘shear strain’. The usual relation between the two would be written:

$$\sigma_{shear} = S\epsilon_{shear} \qquad \frac{F}{A} = S\left(\frac{\Delta x}{L}\right)$$

For every material there is a constant associated with the response to this stress. For shear stresses this is called the shear modulus and usually denoted S . Shear stress tries to make one layer of a material slide over another. Sometimes this is much easier than squashing two layers of a material together. As a result, the shear modulus and the Young’s modulus can be very different. To know how a material will respond, you need to know what kind of stress is applied to it.

A third kind of stress and strain is called bulk stress and strain, or “hydrostatic” stress and strain.



$$\text{Stress} = \frac{F}{A} = P_{\text{hydrostatic}} \qquad \text{and} \qquad \text{Strain} = \frac{\Delta V}{V}$$

This is the kind of stress and strain encountered when the object is under pressure which squeezes in from every direction, like when it is deep under water. For this we write:

$$\sigma_{bulk} = -B\epsilon_{bulk} \qquad \text{or} \qquad \frac{F}{A} = -B\left(\frac{\Delta V}{V}\right)$$

For every material there is a constant associated with the response to this stress as well. For bulk stresses this is called the bulk modulus and usually denoted B . Notice that in the equation for bulk stress and strain we include a minus sign. This is because hydrostatic stress is different in nature from the other stresses we have considered: it acts in every direction! For tensile and shear stress, the force creating the stress acts in a particular direction, and the strain is in that direction. Bulk stress is different, it always pushes inward, and when it is positive, it causes a decrease in volume. So in this case, we include a minus sign, so that a positive pressure (stress) produces a decrease in volume (strain). Sometimes the relation

between bulk stress and bulk strain is characterized in terms of the “compressibility”, defined as the inverse of the bulk modulus of the material.

You can see how closely related these three stress/strain relations are. They’re all just expressions of the basic model of a solid as a collection of springs connecting atoms.

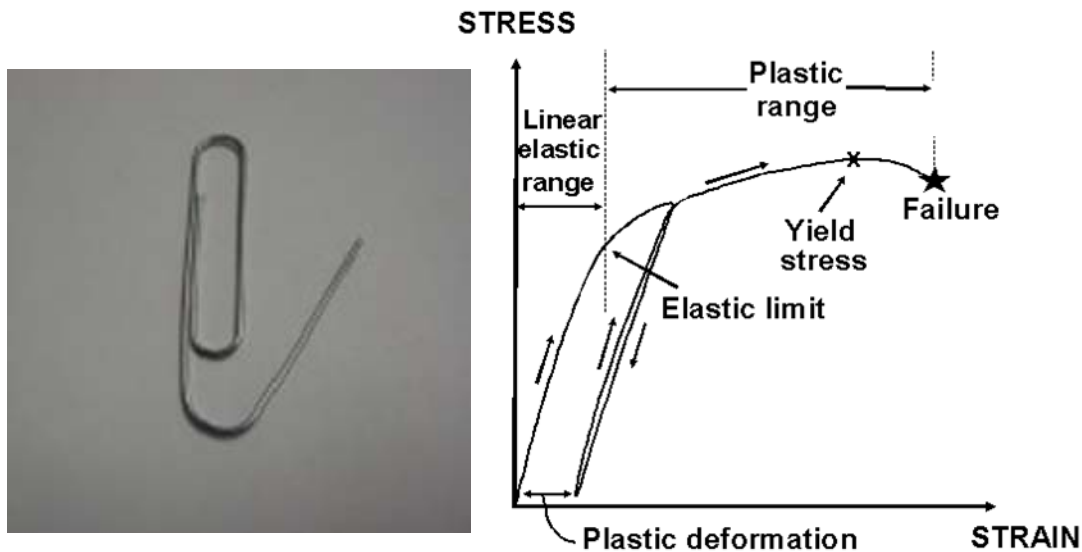
4.6 Limitations to the “Hooke’s law” model of linear stress/strain relations

Hooke’s law is a **linear** model of how stress creates strain. It is an empirical, phenomenological, law. Many materials behave like this under modest stresses. We certainly expect it to break down eventually. It can’t be right when the stress becomes large enough to break the object. Likewise, we know that there are materials with more exotic behaviors. Let’s talk about these limitations in turn.

Strength: plastic deformation and rupture

What happens if the stress becomes too large, and you stretch the object too much?

First, it stops behaving elastically, and no longer returns to its original shape when you remove the stress. For each material, this happens at a stress which is called the ‘elastic limit’ of the material. If the stress rises above this, the material will become ‘plastic’ and permanently deform. This behavior might be familiar to you if you think about bending something made of metal, like a paper clip. Bend it a little, like when you use it normally, and it springs back to its original shape. Bend it a lot, and it stays permanently deformed. This plastic deformation can, of course, be very useful, allowing us to make new shapes using materials which are ductile, like metal and clay. The word ductile describes things which can be plastically deformed without breaking. It is the opposite of brittle, which describes things that break before they will plastically deform.



Eventually the material breaks. We sometimes talk about this happening at the ‘breaking stress’ of the material. But what actually happens when things break is *much* more complicated. We will look at some features of tearing and shattering after we discuss energy. Since stress is related to strain in materials, we could talk about either the breaking stress or the breaking strain of a material. Which one we use may depend on the application.

We can divide materials up in several interesting ways based on how they break:

- “Strong” material: high breaking stress (supports big loads)
- “Weak” material: low breaking stress (can’t support big loads)
- “Stiff” material: low breaking strain (can’t be stretched much before breaking, no matter how large that stress is...)
- “Flexible” material: high breaking strain (can be stretched a lot before breaking)

Here are some illustrative examples of various kinds of materials:

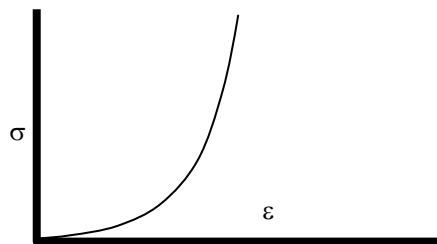
- Steel: stiff, and strong
- Biscuit: stiff, but weak
- Silk fibers: flexible and strong
- Jello: flexible and weak

Nonlinear, non-Hookian flexible materials

Most man-made objects are made of relatively stiff materials that have linear stress/strain relations extending to a large fraction of their breaking stress. They are mostly dry, hard, solids which behave more or less according to Hooke’s law.

By contrast, most biological structures are often made of wet, squishy stuff which doesn’t obey a linear Hooke’s law stress/strain relation. These materials are flexible, able to stretch a lot before breaking, but still very much elastic. They easily deform and then spring back to their original shape. Nonetheless, they are often quite strong, with high breaking strains.

For many of these squishy materials, like your flesh or tendons, the stress vs. strain curve takes a shape called the J curve:



What does this ‘J curve’ stress-strain relation imply?

- Small stress (σ) creates a large strain (ϵ) in the beginning
- Gradually, the additional stress required to increase the strain becomes larger and larger. The “stiffness” increases as the strain becomes larger. This might be expressed by saying that the stiffness is proportional to the slope of this curve of the stress vs. strain curve at each value of strain. Remembering that the Young’s modulus is also the slope of the stress vs. strain curve, this makes a lot of sense.

- At large strain, the material becomes very stiff indeed. To stretch it any further you have to increase the stress very substantially.

You can confirm this general kind of behavior for one biological material rather easily; your earlobe. Try tugging on it. At first, it stretches a lot even when you pull just a little. Eventually though, you reach the steep part of the J curve, and you have to pull a *lot* harder to get it to stretch even a little bit further.

Why does nature use materials like this? What are the advantages?

- The easy ability to reach large strains eliminates the requirement for very fine tolerances in construction. To pick something up, you don't need a hand of exactly the right size. A small force allows your hand to distort, fitting snugly whatever you try to pick up. Try picking up a can of soda with a pair of metal tongs and you'll quickly see this advantage.
- Such materials are robust against shattering; imagine a window which could 'give' and freely change shape when hit by a baseball. Such a window would be able to spread the force of impact over a large area and hence wouldn't shatter.



Examples of times when it pays to be easily flexible: an octopus squeezing itself into a protective shell borrowed from a clam, and a bunch of kids squeezing through a fence to get into a hockey rink.

It is possible that this kind of 'organic' design will expand its influence in human design. For example, perhaps one day cars will become squishy, protective things. Certainly flexible plastic bumpers and air bags are steps in this direction. Steven Vogel, a Professor of Engineering at Duke University, has written a charming book about the relation between human design and evolved form. Called "Cat's Paws and Catapults"², it compares fundamental elements of human and natural technology, noting that:

"Natural and human technologies differ extensively and pervasively. We build dry and stiff structures; nature mostly makes hers wet and flexible. We build of metals; nature never does. Our hinges mainly slide; hers mostly bend. We do wonders with wheels and rotary motion; nature makes fully competent boats, aircraft, and terrestrial vehicles that lack them entirely. Our engines expand or spin; hers contract or slide. We fabricate large devices directly; nature's large things are cunning proliferations of tiny components."

If you're interested in learning more about the differences in these two technologies, Vogel's books on the topic are a great place to start.

A particularly dramatic, and perhaps disgusting, example of the flexibility of biological materials is seen in ticks. Since blood is mostly water, blood feeders like ticks must be prepared to substantially expand their bodies when they eat. Very often, these arthropods eat ten times their own mass in a single meal, quite literally blowing up like a balloon. This picture shows two ticks of the same species, *Amblyomma hebraeum*. The small one is the normal form, prior to a meal, while the large balloon the first stands on is the second, fully engorged tick.



A Quick Summary of Some Important Relations

Conditions for equilibrium:

$$\Sigma \vec{F} = 0 \quad \text{and} \quad \Sigma \vec{\tau}_{\text{every center}} = 0$$

Weight and center of gravity:

The pull of the Earth on an object, its weight, can always be taken to act at its center of gravity, which for our purposes is its center of mass; the mass-weighted average position of the object.

Hooke's law for objects:

Objects respond to loads by yielding to them. For many objects the amount of distortion is proportional to the load, and is well described by Hooke's law:

$$F_{\text{applied to object}} = k\Delta x \quad F_{\text{object pushing back}} = -k\Delta x$$

Stress and strain for materials:

Within an object, we measure stress and strain, rather than force and displacement. There are three kinds of stress and strain discussed here: tensile (or compressive), shear, and bulk. They are defined by the following relations:

$$\begin{aligned} \sigma_{\text{tensile}} &= \frac{F}{A_{\text{along } F}} & \sigma_{\text{shear}} &= \frac{F}{A_{\perp \text{ to } F}} & \sigma_{\text{bulk}} &= P_{\text{hydrostatic}} \\ \epsilon_{\text{tensile}} &= \frac{\Delta L}{L} & \epsilon_{\text{shear}} &= \frac{\Delta x}{L} & \epsilon_{\text{bulk}} &= \frac{\Delta V}{V} \end{aligned}$$

In engineering materials, stress is often linearly proportional to strain, and in each material the constant of proportionality is called the 'modulus' of the material; either the tensile (or Young's) modulus, the shear modulus, or the bulk modulus. In most biological materials, stress is *not* proportional to strain, and a stress-strain curve is required to describe how the material responds when stress is applied.

¹ Labarbera, Michael, "The Strange Laboratory of Dr. Labarbara", University of Chicago Magazine, Oct-Dec, 1996

² Vogel, Steven, "Cat's Paws and Catapults: Mechanical Worlds of Nature and People", W.W.Norton, 1998.

5. Getting around: friction and motion

- 1) Different kinds of force 'laws': phenomenological models vs. fundamental forces
- 2) Resisting relative motion: how friction acts
 - i. Da Vinci and two tendencies of solid friction
 - ii. A linear model: the coefficient of friction
- 3) Frictional forces and practical examples of motion
 - i. Sticking vs. slipping: static and kinetic friction
 - ii. Examples of the role of friction: lifting and standing
- 4) Origins of friction: adhesion and surface roughness
 - i. What are solid surfaces like?
 - ii. Stick and slip friction: μ_s and μ_k
 - iii. How large can μ_s be?
 - iv. Why is friction independent of contact area?
- 5) Breaking the 'rules' of friction
 - i. The J-curve and sneakers
 - ii. Rolling motion
- 6) Moving through air and water: fluid friction in two extremes
 - i. Large things moving fast
 - ii. Small things moving slowly
 - iii. Terminal velocity

Physics of the Life Sciences: Chapter 5

5.1 The place of friction in motion

Now that we have learned a bit about statics, about the balances of forces and torques needed to keep objects in place, it's time to delve a little deeper into how some of these individual forces come about. In this chapter we will concentrate on friction, a very important player in the lives of organisms, and in all of our technology. Without it there is no way we could get around at all.

Friction is the tendency of objects in contact to resist relative motion, to stick together. It is the dominant factor in most motion which we observe in our world. Its tendency to bring any moving body to rest is what so reinforced the Aristotalean view that "motion implies a mover". Because of the resistance which friction provides, experience suggests that a continuous force is required to keep something moving. To uncover the real principles of dynamics, Galileo had to imagine a world without friction. In such a world motion could be perpetual, and no force at all would be required to maintain it.

Constant motion is as *natural* a state as rest. Because of this, Newton focused our attention on forces as the cause of *changes* in motion. He also showed that to predict the motion of an object, all you need to do is to understand the forces which act on it. So if we wish to understand the influence of friction on the motion of bodies, we need to understand how to predict the magnitude and direction of the frictional force.

We will start by examining one particular kind of friction, the sliding friction which occurs when two dry, solid surfaces slip across one another. This is the force which causes a book to slow to a halt when you slide it across the table.

Force laws: phenomenological and fundamental

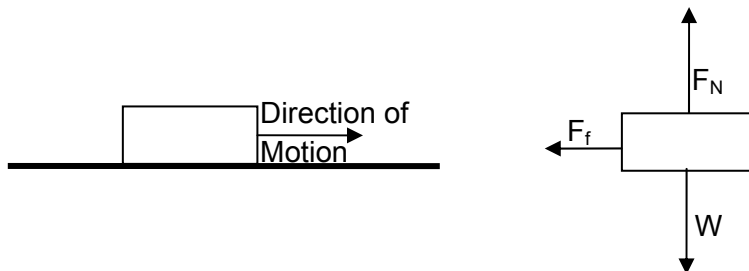
To understand complex forces like friction, we often will seek a ‘force law’. Such a force law is a mathematical model, an equation, which will tell us how large a force on a body would be if it had a particular set of properties (such as mass, composition, surface condition etc). Establishing force laws is a basic task in physics, and the force laws which we derive tend to fall into two categories: fundamental and phenomenological. As we have seen, there are only a few fundamental forces. But these forces often act in complex ways which are better described by approximate phenomenological force laws. We don’t mean to suggest that phenomenological force laws are not accurate reflections of reality or that they aren’t “true”. Such models are sometimes extremely precise. It’s just that by acknowledging that they gloss over details, we confine these models to being “true” with a lower-case *t*. We know for sure that there are other details hiding beneath the general principles these models encode.

To find a force law, we first try to describe the *basic* behavior. We try to predict correctly the approximate size and direction of forces; to understand approximately how these forces will change when we alter the objects in question. We might call this understanding the problem in the "first approximation". Once we have a handle on the basic behavior, we look at things at the next level of detail (in the second approximation), and so on. Phenomenological laws are never perfect, but they can be enormously *useful*, and in complicated cases like friction they are absolutely necessary.

It’s worth remembering here that the structure and behavior of living systems is often extraordinarily complex. As a result, living systems almost always require description with this sort of phenomenological approach. Quantitative models of biological systems almost always begin simple, then gradually add complexity, and accuracy, in this way. Mathematical modeling of this kind is an increasingly important part of life science research. New work in these areas is very vibrant, and is reflected in research programs with names like Mathematical Biology, Biostatistics, and Complex Systems.

5.2 How does friction act?

Friction always acts to resist relative motion between two surfaces which are in contact. Let's consider two examples to see what this means. First the simplest: imagine I slide a block over the table. I push it for a bit and then let it slide to a halt. **While** it is sliding to a halt the friction between the surfaces will generate the force which acts to prevent this motion, slowing the block down until it stops.



In this case, the frictional force acts in a direction opposite to the direction of motion. When friction acts in this way it tends to bring things to rest. This is the kind of frictional effect which led Aristotle to conclude that the natural state of objects was at rest.

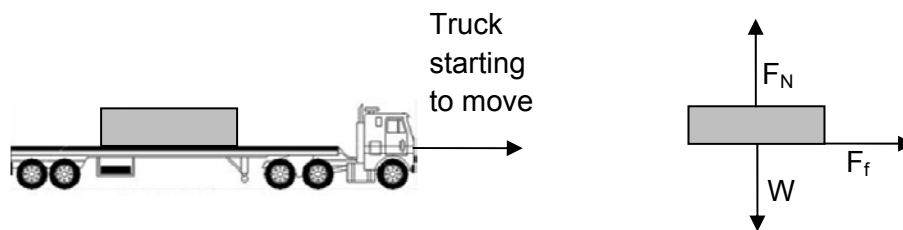
Now let's consider another case. A heavy block sits at rest on a surface. I touch it on the side with my finger and apply a force; but the block doesn't move. It remains at rest. Why?



Since we see the block remain at rest, some force must balance the push I applied, keeping the net force on it zero. This balancing force, which only appears when it's needed to oppose my force, is also a frictional force.

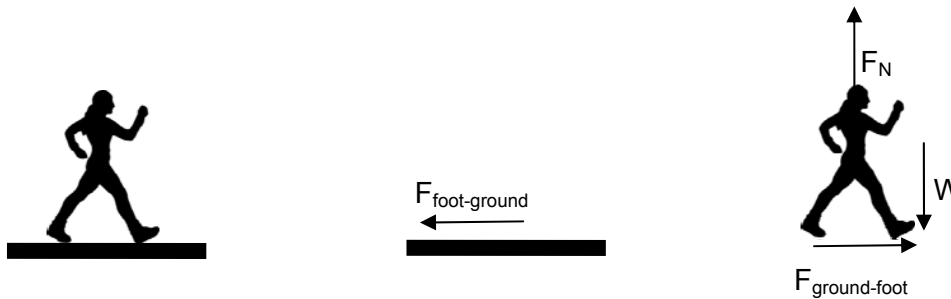
For reasons which we will describe in a minute, we think of these two examples as involving two different kinds of friction. The first case, in which the two objects are already in motion relative to one another, is called "**kinetic**" friction, because it refers to objects which are in motion. Once they are in motion, this kind of friction acts to oppose their relative motion. The second kind of friction, acting before the objects begin to slip, is called "**static**" friction. Static friction acts to prevent objects from beginning to slide over one another.

We generally think of friction as something which prevents motion, but this is selling it short. In fact friction is essential to creating motion, as the following two examples should reveal. Consider our heavy block again, but now imagine that it rests on the bed of a flatbed truck. As the truck begins to move forward, a force is required to move the block along with the truck. The force which does this is a frictional force; it pulls the block forward with the truck in an effort to prevent *relative* motion between the block and the truck. In this case, the frictional force creates the motion of the block. Without it, the truck would be unable to start moving without leaving the block behind.



Consider too how it is that you start to walk. When you stand at rest, there is no frictional force between your feet and the ground. It acts to prevent relative motion. Nothing is trying to create relative motion

when you stand there, so no frictional forces act. If you want to start walking forward, you begin by pushing backward on the floor with your foot. When you push backward on the floor with your foot, the floor pushes forward on you. The interaction between your foot and the floor is a frictional force; it is resisting relative motion (slipping) between your foot and the floor. The force which allows you to start moving forward is the forward frictional force the floor applies to you in response to you pushing backward on it with your foot. Without this frictional force, it would be impossible for you to begin moving forward, as I'm sure you know if you have ever tried walking on wet ice.



Notice that in both of these cases (starting the truck and walking) the friction which acts is static. It is acting to prevent relative motion, between the block and the truck and between the foot and the ground, and in both cases the frictional force is large enough to succeed in preventing relative motion.

So friction is really omnipresent and essential. Without it everything which began moving would continue moving forever; the world would be a crazy jumble like the atoms in a gas. Everything about our ability to start and stop, move where we like, hold a pencil, relax in a chair; it all ultimately comes from friction.

Two tendencies of sliding friction:

The size of frictional forces between two dry solids can be modeled reasonably well using two principles first uncovered by Leonardo da Vinci. Leonardo was an eclectic genius, interested in both the practical and aesthetic arts. He studied friction for a typically complex mix of practical and aesthetic reasons.

During his time, it was thought that the planets were held in their orbits by a set of solid concentric spheres, each of which rotated at a different rate. From classical times it was thought that the true scholar, understanding the beauty of this arrangement, would hear the "music of the spheres", a kind of divine symphony. Leonardo thought perhaps this music was generated by friction between these spheres as they moved relative to one another. His attempts to understand worldly friction were driven, in part, by his desire to experience this divine music.

Da Vinci's two principals are simple rules about the forces which resist the sliding motion of two solid bodies over one another:

1. The force of friction resisting the relative motion of two bodies is directly *proportional* to the normal force between the two bodies.
2. The force of friction resisting the relative motion of two bodies is *independent of* their area of contact and the rate of motion.

The first assertion sounds reasonable; the harder you squeeze two objects together, the more difficult it is to slide them over one another. The second is somewhat surprising; we will see a bit later why it is the case. Remember that these two rules are nothing like absolute laws; they're just general principles that capture the main features of friction in some cases, especially for two dry flat surfaces.

The coefficient of friction

These two rules for friction allow us to construct a mathematical model for how large the friction force is. We're hoping to write an equation which says:

$$F_{friction} = ?$$

The first rule says that $F_{friction}$ is directly proportional to the size of the normal force between the two objects. The second rule says that $F_{friction}$ does not depend on the area of contact or on the rate of motion. So we know the area of contact and rate of motion won't appear in the equation. This allows us to write:

$$F_{friction} \propto F_N$$

We know $F_{friction}$ is proportional to F_N , but what determines the proportionality constant? If this constant is large, there will be a lot of friction between the two objects. If it is small, there will be little. The magnitude of this proportionality constant depends on the properties of each of the surfaces. For surfaces made of any pair of materials, like steel and wood, a single proportionality constant is usually enough:

$$F_{friction} = \mu_{steel-wood} F_N$$

Where $\mu_{steel-wood}$ is the "coefficient of friction" which applies to an interface between steel and wood. Note that since this coefficient relates two forces it is unit-less; a pure number.

This coefficient of friction depends mostly just on the composition of the two surfaces which are in contact. Here are a few examples:

Brick on wood: $\mu_{brick-wood} \approx 0.5$

Ice on Ice: $\mu_{ice-ice} \approx 0.025$

Copper on copper: $\mu_{copper-copper} \approx 0.9$

The coefficient of friction is a measure of how freely surfaces of two materials stick to one another. When the coefficient is high, the surfaces stick together quite a lot. When it is small, they stick together very little. Using the example of brick on wood, we would find:

$$F_{friction} = \mu_{brick-wood} F_N \cong 0.5 F_N$$

So in this case, the size of the force of friction would be about half the size of the normal force pushing the two materials together.

5.3 Frictional forces and practical examples of motion:

To understand how friction acts, it's useful to start with a simple case where the motion is on a horizontal surface. Imagine you place a book on a table and push on it with a horizontal force. If the force is too small, friction will hold the book in place. When you push very lightly, there is a small frictional force holding it in place, because only a small force is needed. When you push harder, the frictional force becomes larger. Eventually, when your force becomes large enough, the book will break loose and start to slide. Once this happens, a force is still required to keep the book sliding, but you will usually find that the force required to keep it moving is less than that required to get it started. Try this yourself and you should be able to get a sense of how it works.

What is there to notice here? First, you can see the difference between static vs. kinetic friction. Static friction is a passive force; it adjusts its value to be just what it has to be to keep the object from moving. As you increase the force with which you push on the book, the static frictional force increases just in step with to prevent the book from moving. It will keep increasing up to some maximum point, at which the book breaks free and begins to move.

We generally find that this maximum static frictional force depends on the nature of the surfaces in contact, together with the normal force, in just the manner described above. That is:

$$F_{friction}^{static-max} \cong \mu_s^{these\ surfaces} F_N$$

In this equation, μ_s is the "coefficient of static friction", and F_N is the normal force preventing the book from moving through the table. Remember that this equation describes the *maximum* static friction force. Because static friction is passive, the actual $F_{friction}^{static}$ is *whatever it has to be* to prevent the object from sliding. It can take any value from zero to the maximum value given by the above equation. Be very careful about this distinction when you attempt to understand static friction problems.

Once we apply a force large enough to overcome static friction and make the book begin to slide, kinetic friction becomes the relevant force. Kinetic friction is better considered an active force than a passive one. Its magnitude is determined all the time by the equation:

$$F_{friction}^{kinetic} \cong \mu_k^{these\ surfaces} F_N$$

In this equation μ_k is the "coefficient of kinetic friction" and F_N is again the normal force between the two objects. So once the book is moving, the frictional force becomes independent of the size of the force you apply and independent of the rate of motion. It depends only on the size of the normal force.

Let's think about this for a moment. You apply a force to get an object moving. If you start out with a small force and gradually increase it, the object will first remain at rest, as the static frictional force gradually increases to match your push. Then, when your push exceeds $\mu_s F_N$, the object will begin to

move. Once it is moving, regardless of the force with which you push, the friction force which resists motion will always be $\mu_k F_N$. Since this frictional force is now *constant*, different things will happen depending on how large the other applied forces are.

If, once it starts moving, you apply a force less than $F_{friction}^{kinetic}$ (perhaps by no longer pushing) the unbalanced frictional force will decelerate the object and slow it down. If you apply a force larger than $F_{friction}^{kinetic}$, the unbalanced part of your force will accelerate the object forward. Only if you apply a force exactly to $F_{friction}^{kinetic}$, no more and no less, will the object move along the surface at constant speed, because only then will the net force along the surface be zero.

This is just what happens when you try to push something heavy (let's say a cabinet) across the floor. When you first push, it goes nowhere. Then you shove harder, and eventually it breaks loose and starts to slip. Once this happens you adjust your force (not too hard, not too weak) so that it slips along at a constant rate. The rate of motion now is not set by how hard you push (you're always just balancing the friction, which is independent of how fast the cabinet moves) but rather by how quickly you're prepared to move along with it.

Coefficients of friction: static and kinetic

The constants in these equations, the coefficients μ , determine the sizes of frictional forces. Their values depend primarily on the composition of the two surfaces which are placed in contact. In detail, they depend on many other things, such as how the surfaces are prepared (are they rough or smooth), and the temperature of the surfaces. But for starters we will stick with the most important factor, what the surfaces are made of. Some examples:

Materials	μ_s	μ_k
Steel on Steel	0.6	0.4
Rope on wood	0.5	0.3
Tires on dry concrete	1.0	0.75
Tires on icy concrete	0.3	0.02
Teflon on teflon	0.04	0.04

A larger table can be found here:

http://www.roymech.co.uk/Useful_Tables/Tribology/co_of_frict.htm#coef

and in many other places online.

There are several patterns among these coefficients worth noting. First, the static coefficient μ_s is larger than the kinetic coefficient μ_k for almost every set of materials. It is harder to start something sliding than

to keep it moving. If you have ever tried to move furniture you will have experienced this. You push hard trying to get something sliding; once it breaks free, it slips along more easily.

The one exception to this general rule shown here is Teflon, for which $\mu_s \approx \mu_k$. Because of this, it is as easy to start sliding motion with Teflon as it is to maintain it, there is no sudden jerk as the two surfaces break free. This property makes Teflon on Teflon contacts very useful for artificial joints, like knee repair. Try moving some of your joints, bending your elbow for example. There is no obvious need to apply an extra large force to get the motion started. This is because your joints, unlike the dry interfaces discussed here, are nicely lubricated.

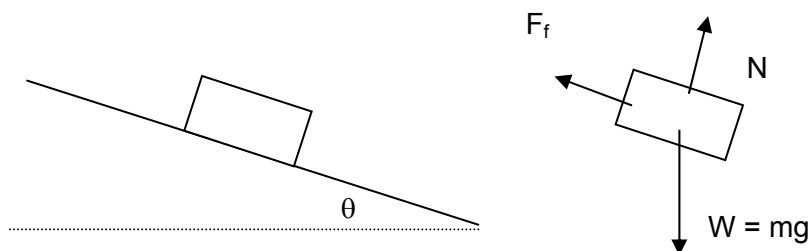
Second, there is a relatively large range of coefficients of friction; they vary from 0.02 to 1.0, another hint that what's really going on here is complicated. And remember, all we're talking about here is the contact between clean, dry surfaces. Imagine how many more different coefficients we would have to know if we wanted to predict the frictional forces for surfaces with varying degrees of contamination, or surfaces which are wet or otherwise lubricated.

This is what we mean when we say these simple laws of friction are a "first approximation" and "phenomenological". They are not a wild guess; real objects do behave in roughly this way. But such phenomenological laws are only capable of giving us a ballpark idea of what's going on. We must use them with caution. What are the essential points? Friction between two dry solid surfaces is approximately proportional to the size of the normal force between the surfaces. The relevant constant of proportionality depends on the nature of the two surfaces in contact. Also, the size of a frictional force can vary when there is no relative motion, but is approximately fixed in size when the two surfaces are slipping over one another. With these basics in hand, let's consider some examples of how friction acts.

Examples of the role of friction:

Let's consider two typical examples of friction. The first involves a book slipping down a slope. Imagine a book sitting on a board, arranged so that we could gradually increase the angle the board makes with the horizontal. What happens?

As we increase the angle θ , the force with which gravity pulls the object down the slope gradually increases. For a while, the static frictional force increases in step with this, balancing the downward pull of gravity so that the block will remain in place. But at some point, the pull of gravity down the slope overcomes the maximum possible static frictional force, and the block slips.



Now let's analyze this situation. To begin, we draw the situation when the board is tilted at some particular angle θ . Then we draw a free body diagram for the book, showing the three forces which act on it. Given this FBD, we should sum the forces along different directions. In this case, the block is free to move only along the slope, and can't move at all perpendicular to it. So we will sum the forces along and perpendicular to the slope.

Before the block begins to slide we have:

$$\Sigma F_{\text{along slope}} = mg \sin(\theta) - F_{\text{friction}}^{\text{static}} = 0 \quad \text{or} \quad F_{\text{friction}}^{\text{static}} = mg \sin(\theta)$$

$$\Sigma F_{\text{perpendicular to slope}} = F_N - mg \cos(\theta) = 0 \quad \text{or} \quad F_N = mg \cos(\theta)$$

But remember, there is a maximum static frictional force, it cannot be larger than $\mu_s^{\text{book-board}} F_N$. Since the frictional force required to keep the book in places increases from zero when $\theta = 0$ continuously until it slips, we can use this information to calculate the angle θ_{slip} at which it will start to slip. This will happen when:

$$F_{\text{friction}}^{\text{static}} = \mu_s^{\text{book-board}} F_N = mg \sin(\theta_{\text{slip}})$$

And since we know the size of the normal force, we can write:

$$\mu_s^{\text{book-board}} F_N = \mu_s^{\text{book-board}} mg \cos(\theta_{\text{slip}}) = mg \sin(\theta_{\text{slip}}) \quad \text{or} \quad \mu_s^{\text{book-board}} = \tan(\theta_{\text{slip}})$$

And find that the angle where it will slip is:

$$\theta_{\text{slip}} = \arctan(\mu_s^{\text{book-board}})$$

Notice what's going on here. There are **two** effects. First the frictional force required to hold it on the slope is **increasing** as we increase the angle θ . Second, the maximum available static friction is **decreasing** as we increase θ , because the normal force between the surfaces is decreasing. Both these facts tend to make objects slip down slopes more easily.

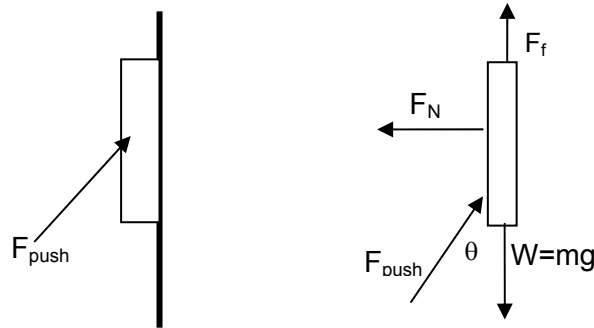
This application of the laws of friction tells us that for a particular coefficient of friction, there exists some maximum slope beyond which the object will slip. Surely this is familiar to you from standing on slopes. If the slope is too steep, you slip downward. You probably have also experienced the important dependence of this critical slope on the nature of the two materials. Stand on a slope in sneakers and you can avoid slipping to quite steep angles. Do it in dress shoes (or on ice!) and the slope you can manage is much less steep.

This result also suggests that the critical angle at which things will slip is independent of the size and shape of the object on the slope. If you stand on a slope with a two year old, you'll both start to slip at the same angle. This might be somewhat surprising, but it's a clear prediction of this result.

Going back to the original example, once static friction is overcome and the block starts slipping, then the friction becomes purely kinetic. Once that happens, we know exactly how large it is:

$$F_{friction}^{kinetic} = \mu_k^{book-board} F_N = \mu_k^{book-board} mg \cos(\theta)$$

Here is a second example for you to think about:



When you push a book against the wall with a force like this, will it slip up the wall, down the wall, or remain in place? The only way to be sure is to work out the details. Since we don't (until we work it out) know what's going to happen, let's begin by imagining that the frictional force the wall applies to the book acts upward. If we find it is different, the answer we get for it will contain a minus sign.

First we sum the forces along and perpendicular to the wall. We will call the direction along the wall the y-direction, with up as positive, and the direction perpendicular to the wall the x-direction, with positive to the right.

$$\Sigma F_y = F_{push} \cos(\theta) + F_f - mg = 0 \quad \text{or} \quad F_f = mg - F_{push} \cos(\theta)$$

$$\Sigma F_x = F_{push} \sin(\theta) - F_N = 0 \quad \text{or} \quad F_N = F_{push} \sin(\theta)$$

OK, so far so good. Now we can apply what we know about static friction to write:

$$F_{static\ friction}^{maximum} = \mu_s^{book-wall} F_N = \mu_s^{book-wall} F_{push} \sin(\theta)$$

Let's think about the limits. If F_{push} is small, then friction must help to support the book, to keep it from sliding down. This is the case we guessed and drew in the diagram above, in which the frictional force acts upward. There is some limit, some minimum force F needed to keep the book from slipping down the wall. This limit is reached when:

$$F_{static\ friction} = mg - F_{push} \cos(\theta) = \mu_s^{book-wall} F_{push} \sin(\theta)$$

Or when

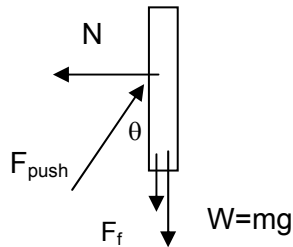
$$F_{push} = \frac{mg}{\mu_s^{book-wall} \sin(\theta) + \cos(\theta)}$$

There is also a particular combination of angle and force for which the frictional force required is zero. This happens when:

$$F_f = mg - F_{push} \cos(\theta) = 0 \quad \text{or} \quad F_{push} = \frac{mg}{\cos(\theta)}$$

Finally, if we push up harder and harder, we are trying to make the book slip **up** the wall. When this happens, we can redraw the picture and rewrite the equations as:

$$F_{push} \cos(\theta) > mg \quad \text{so} \quad F_f = mg - F_{push} \cos(\theta) < 0$$



In this case, the friction between the wall and the book is actually acting downward, rather than up. It is trying to prevent the book from sliding *up* the wall. There is another limit to check in this case. If we push too hard, friction cannot resist our pushing, and the book will begin to slip upward. This occurs when:

$$F_{static\ friction}^{maximum} = \mu_s^{book-wall} F_N = \mu_s^{book-wall} F_{push} \sin(\theta)$$

Putting this back into the equation for motion up and down the wall, we find how much friction is required to prevent the book from slipping upward:

$$\Sigma F_y = F_{push} \cos(\theta) - F_f - mg = 0 \quad \text{or} \quad F_f = F_{push} \cos(\theta) - mg$$

Comparing this to the maximum available static friction, we obtain a condition on F_{push} :

$$\mu_s^{book-wall} F_{push} \sin(\theta) = F_{push} \cos(\theta) - mg \quad \text{or} \quad F_{push} = \frac{mg}{\left[\cos(\theta) - \mu_s^{book-wall} \sin(\theta) \right]}$$

Notice what I did with the signs! We have postulated the situation where the friction is holding the book **down**, rather than up. So I have redrawn it with the force down, and added forces appropriately.

Now this is a subtle, interesting relation. It says that the force required to make the book start slipping upward is given by:

$$F_{push} = \frac{mg}{\left[\cos(\theta) - \mu_s^{book-wall} \sin(\theta) \right]}$$

This tells us some interesting things. What force is required if the denominator of this expression is zero? In that case an infinite force would be required. This happens when:

$$\cos(\theta) - \mu_s^{book-wall} \sin(\theta) = 0 \quad \text{or when} \quad \mu_s^{book-wall} = \frac{\cos(\theta)}{\sin(\theta)} = \cot(\theta)$$

So for an angle θ of 60° , it would be impossible to make the book slide up the wall if the coefficient of friction is more than $\mu_s = 0.58$, and if the angle θ is 85° , the required $\mu_s = 0.08$.

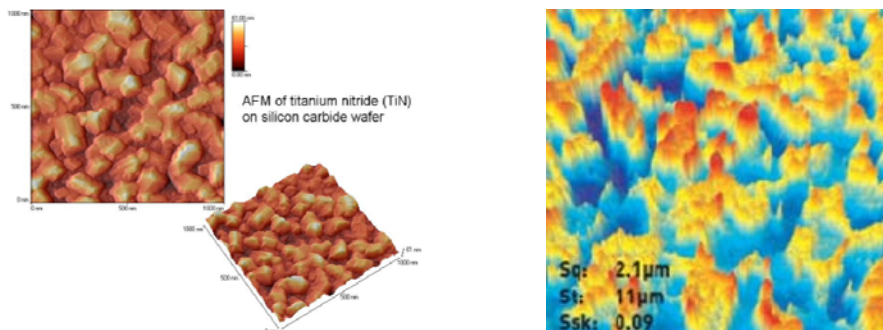
What's interesting about this result is that it isn't completely obvious. Without this careful analysis we might never have guessed that it is *impossible* to make an object slide up the wall under certain circumstances, and we almost certainly would not have guessed that the limiting condition would be independent of the mass of the book we're trying to slide.

5.4 The origins of friction: surface roughness and adhesion

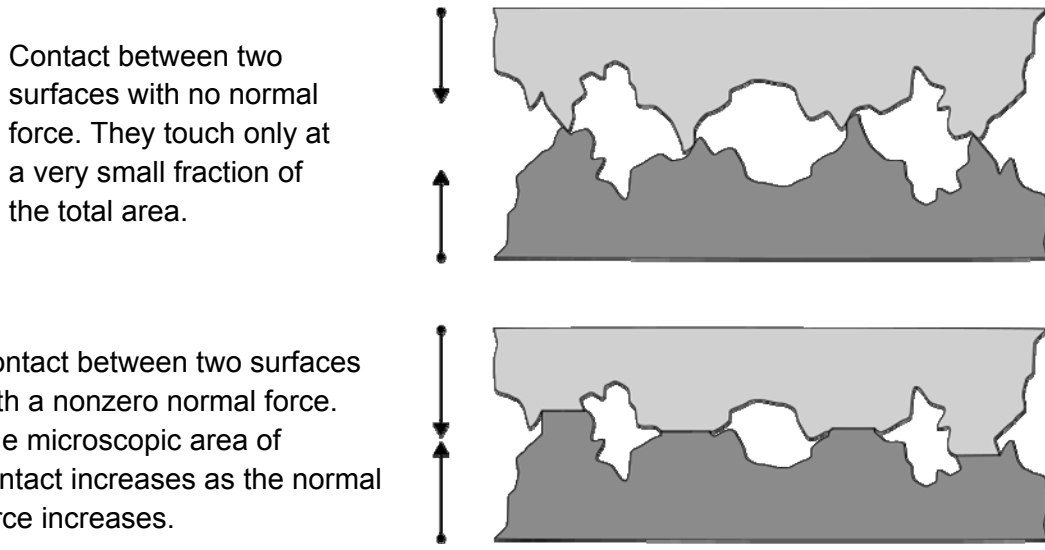
The fundamental origin of this rich, important, and complicated force is of great intellectual and practical interest, and we provide the briefest introduction to it here. To understand friction, you have to begin with an idea of what the surfaces of ordinary objects look like. Are typical objects *smooth* or *rough*? The answer depends on how you look at them.

Is the Earth smooth? You might say no. After all, the largest features on the Earth are about 9 km high. But the radius of the Earth is about 6.4×10^6 m, so the Earth is smooth to about 0.1% of its radius. This is far more smooth, in this fractional sense, than a typical glass marble.

Objects like books, marbles, and tables appear smooth on the macroscopic scale. But when we examine them on an atomic scale, they are very rough, with peaks and valleys that are often many 100s of atoms tall. We can do this now (look at surfaces on the atomic scale), especially using instruments like Scanning-Tunneling Microscopes (STM) and Atomic Force Microscopes (AFM). What we find when we image typical surfaces is something that looks more like the Alps than Kansas.



So when we put two surfaces in contact, it's rather like taking the Alps and turning them over on top of the Himalayas. In such a case, the actual microscopic area of contact is a tiny fraction of the total macroscopic area of the two objects. At the relatively few points where the objects *do* meet, the atoms from the two materials may actually bond, and the materials stick to each other (or "adhere"). This happens through the same kind of chemical bonds which hold the object themselves together. The "sticking" of these points is what we perceive as friction.



It may be surprising to you that two solid surfaces in contact might bond together. After all, when you place one piece of metal on another, they don't typically stick together. This fact, that they don't really stick, is a consequence of the complicated structures of surfaces. If two surfaces are truly flat and clean, on the atomic scale, they may bond together very strongly when placed in contact, so strongly that the interface between the two pieces will become just another layer within the material. This is used sometimes in a manufacturing process called cold welding.

With this picture in mind, we can start to understand da Vinci's two rules and some other features of friction.

The laws of friction and independence of area

Ultimately all friction is caused by this bonding between atoms, and all such bonds are due to electromagnetic interactions between the atoms. The same interactions that hold matter together create the stickiness which underlies friction. How large this effect is depends on both the materials you use and the nature of the surfaces (polished so that many atoms come into contact, or Alps-on-Himalayas so that very few come into contact). But since most surfaces are actually quite rough, the simple laws da Vinci first discovered give a pretty accurate estimate of what will happen.

Probably the most surprising thing about sliding friction is that it doesn't depend on contact area. How can this be, particularly in light of the fact that friction is really due to adhesion? The trick to this, not understood until at least the 1950s, is that friction *does* depend on the area of contact, but not on the

apparent area of contact. It depends instead on the tiny bit of contact at the tops of those mountains, what we might write $A_{\text{contact}}^{\text{microscopic}}$. With this knowledge we can write

$$F_{\text{friction}} \propto A_{\text{contact}}^{\text{microscopic}}$$

And for most solids, which are very rough indeed, this microscopic area of contact depends on how hard you squeeze the two surfaces together:

$$A_{\text{contact}}^{\text{microscopic}} \propto F_N$$

Hence:

$$F_{\text{friction}} \propto F_N$$

This is a great example of how going deeper, looking beneath the first level of phenomenological laws, can be revealing. Understanding the reason *why* friction is independent of contact area makes it possible to better appreciate the limitations of this general rule, as we will see in a moment.

Stick-and-slip friction, μ_s , and μ_k

A second interesting question we want to explore is why μ_s and μ_k are different, and why μ_s is typically larger. When I try to move a stationary object I gradually stretch these “merged mountaintops”, causing them to distort and resist the force which I apply. Once the object breaks free and starts to actually move, what happens goes something like this:

- The distorted object breaks free at the surface, releasing the stretched points of contact which spring forward until they "catch up" with the bulk of the object.
- Then the two surfaces are again nearly at rest. New points of contact bond, generating a new source of frictional force; the surface on top grabs hold of the surface below.
- As we continue to pull, these new contacts are stretched, until finally they break loose, allowing the contact points to "jump" again.

This cycle, which is known as "stick-and-slip" friction, is the reason why static and kinetic friction are related in the way they are. Kinetic friction is really a bunch of repeated applications of static friction. In each step the static frictional force builds up from zero to its maximum value, then breaks free and starts again. Each time the bonds between these two surfaces break, the top material jumps forward. After this, you have to build up the force to overcome static friction in this new spot. This makes the *average value* of kinetic friction somewhat less than the maximum for static friction.

If you want to feel this stick-and-slip friction in action, try putting the eraser of your pencil down on the desk, while you push down fairly hard. As you drag it across the table, it will perform just this kind of stick-slip motion. The horrible shriek of chalk on a blackboard is also just this stick-slip happening, now at a high frequency, so it generates a high pitched noise.

How large can μ_s be?

There is, in principle, no limit to the size of μ_s . It often seems that since $F_F = \mu F_N$, that the frictional force can never be larger than F_N . But this is *not* the case. Friction depends on adhesion, so it is possible for the friction between objects to be much larger than F_N . This is just what we use glue to do, to make the adhesion large enough to dramatically increase the maximum static friction. Then objects will not slip over one another.

5.5 Breaking the rules of friction

As we have several times stressed, these rules are basic, but generally work well for dry, solid surfaces. They *don't* work for a lot of practically important cases, especially for ones involving biological materials. Substantial violations of these rules occur when the materials are neither dry, nor rigid. Many biological cases are like this, in your joints for example.

In these lubricated cases, frictional forces can be very different from what is described here, depending on factors like temperature and velocity, instead of just on the nature of the surfaces and the normal force. There is often little difference between static and kinetic friction in these cases, which is why your joints don't experience the same jerky start that happens when you try to slide a cabinet across the floor.

A good technological example, drawn from biology, is provided by rubber shoes and tires. Rubber is a substance which distorts quite easily, following something like the biological J-curve stress vs. strain relation. The fact that it can distort so easily means that I can make the microscopic area of contact between the rubber and a floor very large without pushing down on it too hard. This large area of microscopic contact means a lot of adhesion, which in turn means large friction.

You can see the efficacy of this if you do a little long jump on a tile floor. Rubber shoes can provide the relatively enormous frictional force required to stop you. If instead you place a sheet of paper on the floor, and jump onto that, your feet will stick nicely to the paper, but the paper slide and you'll land on your can. You may have experienced this with hard soled dress shoes as well.

We can see how this might work by constructing a simple model. If we approximate this effect by saying that, for rubber:

$$A_{\text{contact}}^{\text{microscopic}} \propto F_N^2$$

We would expect to find:

$$F_{\text{friction}} \propto F_N^2$$

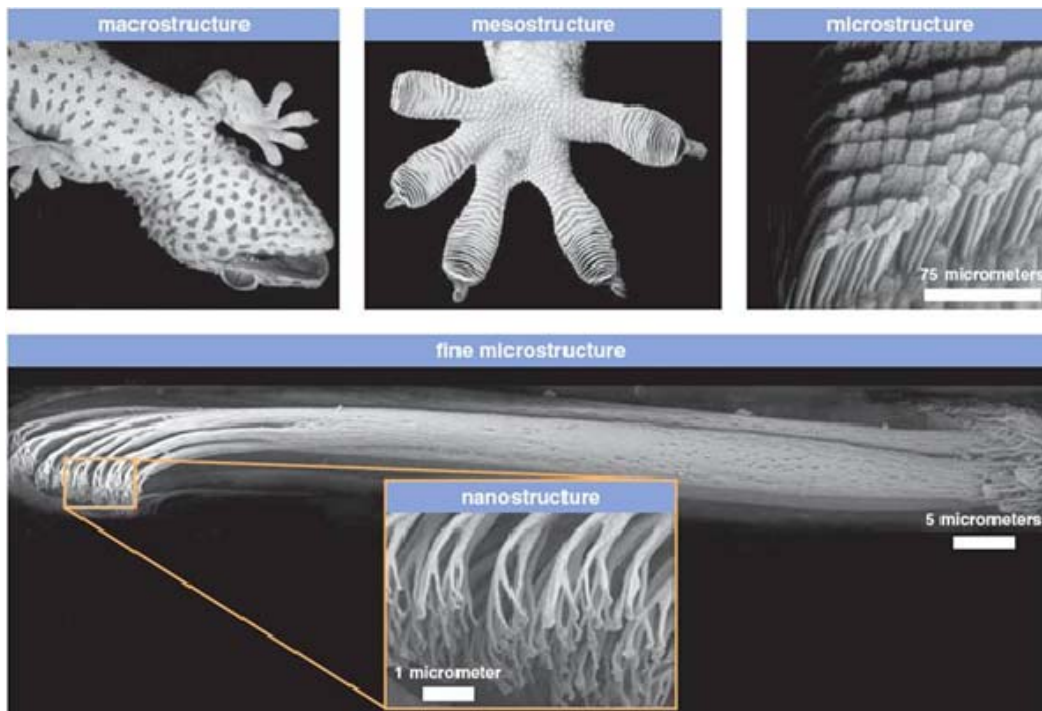
In fact this is approximately what we see for things like rubber tennis shoes.

Other violations of these general rules come from surfaces which are unlike the "typical" surfaces considered here in other ways. It is possible to make extremely smooth and clean surfaces. This is often done in machining to make something called "gauge blocks" out of metal. If I clean two of these gauge blocks carefully and put them together, they will form such a large area of contact that they will

essentially merge into one block in a process known as "cold welding". All the atoms from one surface bond with all the atoms from the other, and the surfaces essentially disappear.

An excellent biological example is provided by the wonderful feet of the Gekkonidae; small, colorful lizards found in warm places around the world. Geckos are famously able to walk straight up walls, and even across ceilings, in nocturnal pursuit of the insects they eat. While this has been known for millennia, (Aristotle marveled at their abilities) the mechanism for this trick was not understood until around 2002.

It is now clear that the impressive adhesion of Gecko feet comes about from a relatively enormous contact area. Each Gecko toe is covered with billions of incredibly finely structured 'setae' which can create an enormous contact area with all kinds of surfaces, smooth or rough. This large contact area provides dry adhesion easily able support the Gecko's weight, allowing them to freely walk up walls and across ceilings. These nanostructures are now being artificially manufactured, and we might expect new kinds of adhesive-free couplings in the future.



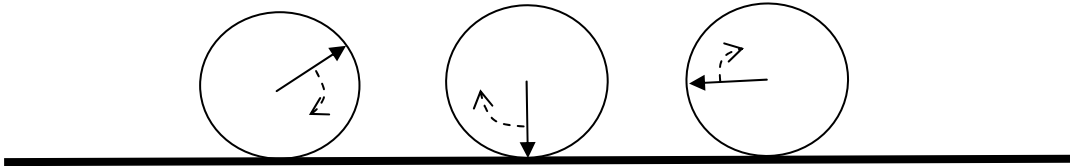
You can learn more about this, and watch some Geckos in action, in this TED lecture:

http://www.ted.com/talks/robert_full_learning_from_the_gecko_s_tail.html

This remarkable approach to adhesion has evolved independently at least a few times, and is also found in the feet of arachnids (jumping spiders) and insects (though these typically add an adhesive fluid as well). It provides another nice example of convergent evolution. Friction is a sticky problem, and the limited variety of good solutions has been repeatedly explored.

Rolling friction

Another practically important example of breaking the ‘rules’ of friction is rolling motion. Why is it that a wheel can roll along for so long, apparently unaffected by friction, while a box that you slide across the same floor so rapidly skids to a halt? The very small friction associated with a wheel comes about because, perhaps surprisingly, the point of contact between the wheel and the ground does not move. Since there is no relative motion between the wheel and the ground, there is no sliding friction. How can this be? Consider the picture below:



In this picture, we see a wheel rolling to the right at three different moments. As it rolls, the spot on the wheel to which the arrow points is set down on the ground, then lifted off again. While it is in contact with the ground, that spot on the wheel is not moving relative to the ground at all. This is why the friction between the wheel and the ground is so small, and consequently why the invention of the wheel is such a big deal. Wheels, balls, and other rolling objects move across surfaces with extremely little friction. Taking advantage of this makes it much easier and cheaper to move things around. Imagine how difficult it would be to get a car from here to Cleveland if you had to overcome sliding friction the whole way!

Rolling motion is a fantastic way to move with little friction. It is important enough in our technology to make it the comparison of choice (“xxx is the greatest invention since the wheel”). Why then was the wheel not discovered by evolution? This question was rather delightfully debated during the 1980s. On one side, paleontologist Stephen Gould argued¹ that the wheel was too complex to evolve. He felt that a rotating joint through which resources would have to be pumped was beyond the limits of what evolution could accomplish. On the other side, anatomist Michael Labarbera pointed out² that wheels aren’t much use unless you have a hard, flat surface to roll on. In the rare natural environments where these conditions exist, like deserts with dry packed soils and small scale regions like leaves, rolling motion is used by organisms as various as dung beetles, spiders, caterpillars, armadillos, and tumbleweeds. Indeed the wheel is only used as the favored mode of transport by people in the right circumstances. Legs remain able to explore a much wider variety of terrain, which is why we still walk around our houses, offices, and classrooms.

5.6 Motion through air and water

When an object moves through a fluid (like air or water) it experiences a force of friction. This is familiar enough if you imagine sticking your hand out the window of a moving car. What is the origin of this fluid friction? As your hand moves through the air, it has to exert a force on the air to move it out of the way. When it exerts this force on the air, the air exerts an equal and opposite force on the hand. This is why airplanes need to run their engines constantly in flight; the force exerted by the engines just balances the force exerted by the air friction.

We would like to construct a mathematical model for fluid friction. Just as with sliding friction, fluid friction is very complicated, and even a simple understanding of it will require considering several possibilities. For fluid friction, there are two useful limits to consider. Which version is most appropriate depends on the circumstances.









One limit occurs when the objects moving are "relatively large", moving "relatively fast", and the fluid flows "relatively freely". In this case the frictional force depends on the size of the object, its speed, and the density of the fluid, in a manner which we can model in the following way:

$$F_{\text{fluid friction}}^{\text{large-fast}} \approx \frac{1}{2} C \rho_{\text{fluid}} A v^2$$

In this equation, ρ_{fluid} is the density of the fluid (in kg/m^3) and v is the speed of the object relative to the fluid. The parameter A is the "cross-sectional" area of the object, the area you would see if the object was moving straight toward you.

The remaining term C is called the "drag coefficient". Its value depends on the shape of the object and the properties of the fluid. For typical objects, neither especially blunt nor elegantly streamlined, this drag coefficient ranges in size from about a half to a bit more than one. So this kind of fluid friction increases linearly with area of the object (more air must be moved out of the way), linearly with the density of the fluid (more mass must be moved out of the way) and as the square of the speed of the object. If an object goes twice as fast, the force which resists it becomes four times as large.

This drag coefficient C is, in a sense, a term we insert to hide all the details. That being so, why is there still a factor of a half in front of this equation? We could simply absorb the factor of a half in the unknown factor C , but the convention is to leave it separate like this. It's worth remembering that the drag coefficient can hide significant additional variation. The fluid friction experienced by a spread-eagled skydiver is quite different from that experienced by someone falling head-first. Examples of drag coefficients for various shapes are shown below.

Shape	Drag Coefficient
Sphere → 	0.47
Halfsphere → 	0.42
Cone → 	0.50
Cube → 	1.05
Angled Cube → 	0.80
Long Cylinder → 	0.82
Short Cylinder → 	1.15
Streamlined Body → 	0.04

Fluid friction is important for organisms when it becomes large compared to other forces which act on them, like their weight. For large organisms living in air, this happens only at quite high speeds, and the obviously streamlined shapes needed to reduce the drag coefficient are restricted to a predatory subset of birds. The large density of water makes fluid friction an important challenge for all aquatic creatures even at relatively low speeds. It's in the water that the dramatic streamlining has repeatedly evolved. Examining the shapes of comparable but unrelated predators like tuna, dolphins, penguins, squid, seals, and ichthyosaurs (an extinct group of marine reptiles) makes this rich case of convergent evolution clear.



The other limit for fluid friction occurs when the objects moving are "relatively small", moving "relatively slowly", through a fluid which flows "relatively poorly". This last part is expressed by something called the "viscosity" of the fluid. Viscosity is a measure of how much internal friction there is in the flow of the fluid. Viscous fluids are things like honey, less viscous materials are things which flow more freely, like air.

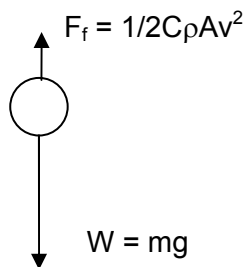
In this case of small, slow things in stickier fluids, the force of friction depends on the size of the object, its speed, and the viscosity of the fluid. If the object is a sphere, the fluid friction will be well approximated by:

$$F_{\text{fluid friction}}^{\text{small-slow}} \approx 6\pi\eta_{\text{fluid}}rv$$

Where η_{fluid} is the viscosity of the fluid, r is a measure of the radius of the object, and v is the speed of the object relative to the fluid. Notice that this kind of friction contains no “drag coefficient” fudge factor. The reason is simple. In this small-slow case, the drag is almost entirely due to friction in the flow of the fluid itself, and not to the inertia of the fluid. The amount of flowing which has to happen as the object moves through the fluid is not very dependent on the detailed shape of the object. There is still room to fudge however. What should this parameter r be? If an object is not round, you can estimate the friction it will experience by choosing a length scale which roughly represents the size of the object when viewed head-on.

The effect of friction on falling objects

How does fluid friction affect motion? Consider what happens to something like a ball falling through the air. Let’s assume this is a sizable ball, so that the ‘large-fast’ form of friction will apply:



If we sum the forces in the y direction:

$$\Sigma F_y = F_{\text{friction}} - W = \frac{1}{2} C \rho A v^2 - mg = \frac{dp}{dt}$$

Initially it’s not moving through the air, $F_f = 0$, and the object accelerates downward with an initial acceleration g . Then, as it picks up speed, the frictional force increases, until eventually it reaches a maximum speed for which the frictional force just balances the weight. Then the net force on the object is zero, and its momentum no longer changes. When this happens, $\frac{dp}{dt} = 0$, so:

$$F_{\text{friction}} - W = \frac{1}{2} C \rho A v^2 - mg = 0$$

The speed at which this happens is called "terminal velocity". We can find it by solving the equation above for the velocity:

$$v_{\text{terminal}} = \sqrt{\frac{2mg}{C \rho_{\text{fluid}} A}}$$

Notice what this tells us. It says that if we increase the area of the object A , v_{terminal} is reduced. If we increase the density ρ , v_{terminal} is reduced. But if we increase the weight, v_{terminal} is increased. More

important, it tells us how rapidly these things change. Double the weight and you get $\sqrt{2}$ times as large a v_{terminal} .

There is a connection here to the scaling laws we have been emphasized at the beginning of this book. If we just take an organism and scale it up in size, its volume will increase like size^3 (increasing the mass), while its cross-sectional area will increase like size^2 . As a result the terminal velocity will increase like $\text{size}^{1/2}$. While this might seem a modest functional dependence, it can be a very large effect. The difference in size between a mouse and a horse is about a factor of 70. This suggests that the terminal velocity of a falling horse would be about 8 times larger than that of a mouse. This is why people say ‘the bigger they are, the harder they fall’. The great early 20th century biologist J.B.S Haldane wrote a delightful essay on scaling laws called “On Being the Right Size” which summed this up nicely:

You can drop a mouse down a thousand-yard mine shaft; and, on arriving at the bottom, it gets a slight shock and walks away, provided that the ground is fairly soft. A rat is killed, a man is broken, a horse splashes.³

If, instead of being large and moving fast, the object was small, traveling slowly, or moving through a more viscous fluid, the friction force law would differ in detail. But an important point is that the motion would be qualitatively identical. When you drop the object, it starts out with a downward acceleration g . Then as it speeds up through the fluid, the frictional force resists its motion, gradually reducing the acceleration until eventually it reaches a stable “terminal velocity” where the frictional force balances the downward pull of gravity.

Using the force law for small objects which we wrote above, we’d find:

$$6\pi\eta_{\text{fluid}}rv_{\text{terminal}} = mg \quad \text{or} \quad v_{\text{terminal}} = \frac{mg}{6\pi\eta_{\text{fluid}}r}$$

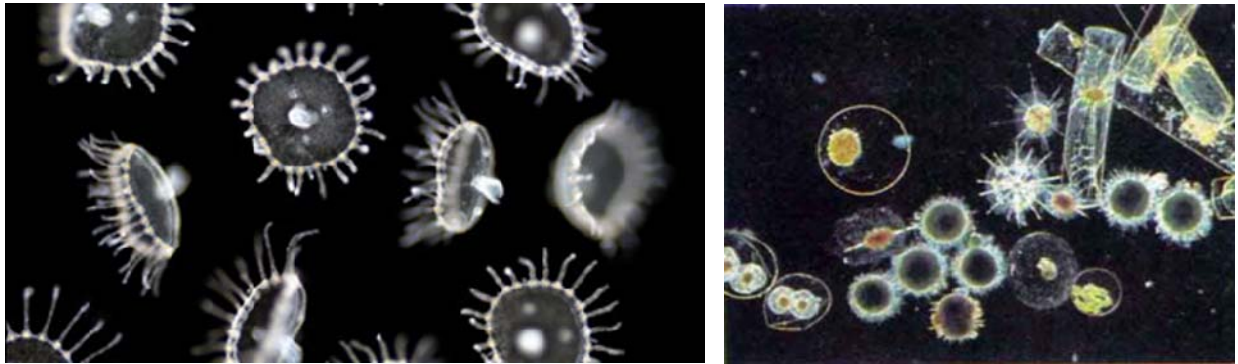
Notice how this is different from what we found above. For these small, slow things in sticky fluids, terminal velocity depends linearly on the mass. Something with twice the mass will reach a terminal velocity twice as large.

Scales, free fall, and life

Life lives largely in fluids, either air or water. The interactions between living things and these fluids play a huge role in life, allowing motion, providing oxygen, and carrying sound and scent to our senses. This example of fluid friction and free fall is just the first encounter we’ll have with this very rich topic. One thing it illustrates at the outset is that the way our fluid environment affects life will be very dependent on scale. The behavior of very small things (individual cells and the world of tiny multicellular creatures) in fluids is very different from that of the world of macroscopic animals.

Just to give one example, there is a class of microscopic things living in the ocean which are called “plankton”. These are organisms so small that the frictional forces applied to them by the water they live in easily overwhelm the downward forces gravity applies to them. Any little flow of water simply carries them with it. In practice, many organisms have evolved means to enhance this, developing shapes which increase fluid drag. This allows them to become somewhat larger while retaining their ability to take

advantage of the free motion riding along with the fluid provides.



Plankton are not *entirely* limited to water. There are things which, for at least part of their life cycle, take advantage of fluid friction in the air to further their ends. Perhaps the most obvious are seeding strategies. Fungi and plants, for the most part, cannot move. If their seeds are to spread they will have to be carried by something else. Quite often, plants enlist mobile animals for this purpose. But many also take advantage of the air, releasing very small seeds or spores, and often equipping them with mechanisms for enhancing the friction between the fluid and the seed (think of the dandelion, cottonwood, or the little maple seed propeller). Another quite beautiful example is the ballooning of baby spiders, which perhaps you will remember from the end of E.B. White's "Charlotte's Web".



Cottonwood seeds

Maple seed

Ballooning spiders

A Quick Summary of Some Important Relations

A model for sliding friction between dry surfaces:

Friction between typical, dry, solid surfaces is generally proportional to the normal force, and can be modeled as:

$$F_{\text{kinetic}}^f = \mu_k F_N \qquad F_{\text{static}}^f \leq \mu_s F_N$$

Keep in mind that kinetic friction is an active force, always the magnitude described here, while static friction is passive, as big as it needs to be, up to the limit described by this equation. The static coefficient of friction μ_s is usually larger than the kinetic coefficient μ_k . Friction ultimately depends on the microscopic area of contact between surfaces. When the microscopic area of contact is not proportional to normal force, as in the feet of geckos, friction will not be well described by the simple model above.

A model for fluid friction:

When objects move through fluids the friction they feel is complex. It may be simply modeled in two extremes: when objects are small and moving slow, and when they are large and moving fast.

$$F_{\text{fluid friction}}^{\text{small-slow}} \approx 6\pi\eta_{\text{fluid}}rv \qquad F_{\text{fluid friction}}^{\text{large-fast}} \approx \frac{1}{2}C\rho_{\text{fluid}}Av^2$$

For small objects at low speeds friction is dominated by the stickiness or ‘viscosity’ of the fluid, represented here by the parameter η_{fluid} . For large objects moving at high speeds friction is dominated by the density of the fluid, represented here by the parameter ρ_{fluid} . In both cases the size of the moving object affects friction, as well as the speed of the object through the fluid.

Terminal velocity:

When the fall of an object under the pull of gravity is resisted by fluid friction it reaches terminal velocity. The value of this depends on the objects mass and shape, as well as the properties of the fluid.

$$v_{\text{terminal}}^{\text{small-slow}} = \frac{mg}{6\pi\eta_{\text{fluid}}r} \qquad v_{\text{terminal}}^{\text{large-fast}} = \sqrt{\frac{2mg}{C\rho_{\text{fluid}}A}}$$

¹ Gould, S., 1981, “Kingdoms without wheels”, *Natural History*, **90**, (3), 42.

² LaBarbera, M., 1983, “Why Wheels Won’t Go”, *The American Naturalist*, **121**, (3), 395.

³ Haldane, J.B.S., “On Being the Right Size and Other Essays”, Oxford University Press, 1985.

6. Describing and quantifying motion in one dimension

- 1) Establishing a basic description
 - i. Position and intervals of distance
 - ii. Instants and intervals of time
 - iii. Motion: position vs. time histories
- 2) Details of the description: changing position
 - i. Changing position and rates of motion
 - ii. Velocity vs. time histories
 - iii. Finding displacement from velocity vs. time histories
- 3) Details of the description: changing velocity
 - i. Changing velocity and acceleration
 - ii. Acceleration vs. time histories
 - iii. Finding change in velocity from acceleration vs. time histories
 - iv. What if accelerations change? Jerk, snap, and beyond...
- 4) Relating these three descriptions of motion: position, velocity, and acceleration
 - i. Three trivial examples: constant position, constant velocity, and constant acceleration
- 5) Motion as a model for all change, and the origins of calculus

Physics for the Life Sciences: Chapter 6

In this chapter we will take a bit of a pause. For a while now we have been examining forces and how they play against one another in cases of equilibrium (either rest or uniform motion). Today we'll begin to discuss how motions **change**, and what happens when forces acting on objects are not balanced. To do this, we must first refine our tools for describing motion.

If we know the full path of an object, its **position** at each instant of time, we know everything about how it has moved. To *understand* this motion, we will need to speak of the **velocity** of the object, how rapidly its position is changing, and its **acceleration**, how rapidly the velocity is changing. For starters, we'll talk just about motion in a straight line. Later we will see that motion in two and three dimensions is a rather straightforward extension of one dimensional motion.

We begin with an example, just to get a sense of where we're headed. Picture in your mind a sprinter prepared to run the 100 m dash. Before the start she is still on the starting blocks. During this time her position remains the same from instant to instant, her speed is zero, and since her speed is not changing her acceleration is zero. Then the gun fires, she bursts forward from the blocks, accelerating quickly toward her top speed. During this period, her position changes from instant to instant. Her speed changes from instant to instant as well, becoming larger and larger. Since her speed is changing, she is accelerating as well.

After just a few seconds, our sprinter is going full out, running at absolutely top speed. During this period, her position continues to change from instant to instant, but her speed does not. Since her speed is not changing her acceleration is now zero.

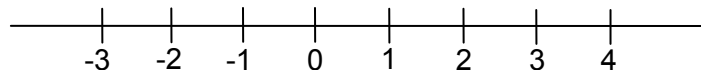
After bursting through the finish tape, our sprinter cruises to a stop. During this period, her position continues to increase, always moving farther from the starting blocks. Now her speed is gradually decreasing, and this changing speed implies an acceleration.

Finally, she stops, hugging her coach in victory. Now her position remains the same from instant to instant, her speed is zero, and since her speed does not change, her acceleration is zero as well. In this chapter, we will develop in more detail the tools we need to describe this motion fully. We will speak of position, velocity, and acceleration, each determined instant-by-instant along the path.

6.1 Position, and intervals of distance

The motion of real, extended objects is complex. You can see this by considering what we mean by a cloud traveling at 10 mph, or a horse racing at 35 mph. When we talk informally about this our meaning is clear (the cloud is 'on average' moving along at 10 mph), and to begin we will simplify complicated motions in this way. We will begin by ignoring the internal structure of an object and treat it instead as a "point object", asking only how this point moves. When we do this we can model a complicated motion (like a horse racing down the track with its legs churning along and its rider bobbing up and down) in a simple way. Once we have a basic model in hand, we can gradually add more and more realistic details.

To describe a motion we begin by setting up a reference frame, a standard scale against which to measure.



To describe motion we simply record the position of the object at every instant. We'll use this kind of labeling scheme.

$$s_1 = \text{position at a particular instant } t_1$$
$$\Delta s_{21} = s_2 - s_1 = \text{interval between the two instants } t_1 \text{ and } t_2$$

Notice the notation, the symbol Δ (the Greek letter "delta") is used to denote a change in a quantity. Think carefully about what the variable s_1 measures. It really only denotes a location. We label this location, this point in space, by noting its distance from the completely arbitrary origin of our coordinate system. You should also think about what the signs of Δs_{21} mean. When it is positive, the object has moved farther to the right between instants t_1 and t_2 . When Δs_{21} is negative, the object has moved to the left.

The signs of the positions and intervals can be confusing, and it may help you to think about what it means to have both s_1 and s_2 be negative, while Δs_{21} is still positive. You might also ponder whether $\Delta s_{21} = 0$ implies that no distance was traveled between instants t_1 and t_2 .

Instants and Intervals of Time:

Time too must be measured. We measure time by comparison to something which happens regularly. What do we compare to? Over the years, many steady timekeepers have been used. The oldest are astronomical, including the rotation of the Earth and its orbit around the Sun. These allow us to mark off days and keep track of the years. To measure shorter periods of time requires something which repeats more often. For this purpose, many different tools have been used, including the pulse, water clocks, masses on springs, pendula, and more recently the very regular, rapid oscillations of quartz crystals and atoms.

In a manner very similar to the way we described positions and intervals of distance, we also talk about instants and intervals of time:

t_1 = time of a particular instant when something happens (an 'event')

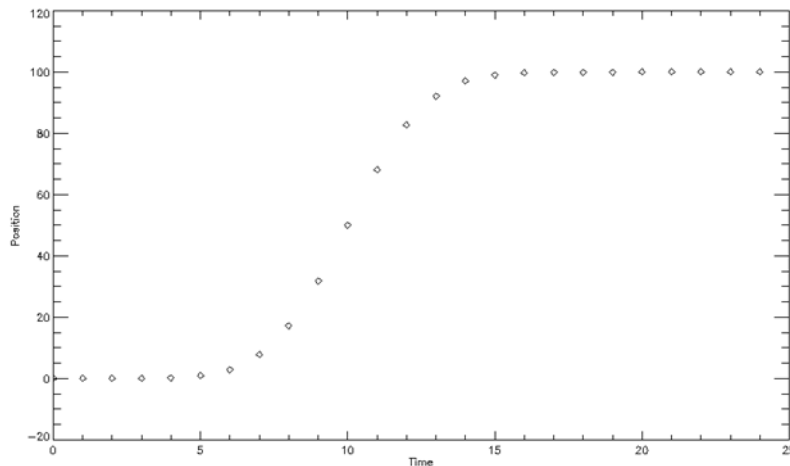
t_2 = time of a second instant when something else happens

$\Delta t_{21} = t_2 - t_1$ = interval of time between two events

Like a position, an instant (t_1 or t_2) is a location in time. It's really just a marker, without units. Only an interval, the time between two instants, has units of seconds.

Motion: position vs. time histories

Since each position s_i corresponds to a particular instant t_i , we can represent the series of events which makes up the motion of an object graphically. This is an example of a position-time graph for a motion:



If the motion is smooth we can reliably fill in the history between these specified points with a continuous curve describing the motion. This curve is a model for actual motion of the object. What does the above picture describe? This object starts at the zero point. It remains there for a bit, not moving as time passes. Then it begins to move to the right, speeding up for a while, and slows to a stop at a positive position of about 100. Interpreting position time graphs properly can be very helpful. In class we will do quite a bit of practicing. Now we want to extend our discussion of motion to include the idea of speed.

6.2 Changing position and rate of motion

When we talk about rate of motion, we want to record how rapidly the position of an object is changing. In everyday life, we would talk about the object's speed. To make things somewhat more precise, we will distinguish between speed and velocity. To fully describe the rate of motion of an object, we have to say both how fast it is moving (its speed), and also what direction it is traveling in. This full description, a vector quantity, is the **velocity** of the object. The speed is just the magnitude of this vector.

Velocity is a good example of something we talk about all the time without care. It clearly has something to do with how fast you go, how much distance you travel in some short period of time, but what's a short period of time? For a cross country trip, it might make sense to consider the entire length of the trip. If we use this in determining velocity we will learn an **average velocity**. If, instead, we wish to know the velocity at an instant, we must examine the distance traveled in an infinitesimally short period. This will tell us the **instantaneous velocity**.

We use the notation for spatial intervals, or displacements, and time intervals to define these:

- Average velocity: Uniform velocity required to travel distance Δs in time Δt

$$v_{average} = \frac{\Delta s}{\Delta t}$$

Average velocity tells you nothing about variations in velocity which might have happened during the interval Δt , but it can give you a good estimate of the object's motion throughout this period of time.

- Instantaneous velocity: the velocity of the object at a particular instant. This can be estimated by examining a time interval Δt which spans the instant t_i at which you want the velocity. To improve the estimate, you continually narrow the time window Δt until it becomes and infinitesimally short dt .

$$v_{instantaneous}(t_i) = \lim_{\Delta t \rightarrow 0} \frac{\Delta s_{\text{around } t_i}}{\Delta t_{\text{around } t_i}} = \left. \frac{ds}{dt} \right|_{t_i}$$

Instantaneous velocity tells you exactly the velocity at this moment, but tells you nothing about velocity at other times.

The average velocity is easy to understand, it's just distance traveled divided by time, but it fails to take into account variations in velocity. In cases where there are no variations, it's fine. The instantaneous velocity is just the derivative of the position with respect to time, ds/dt . This idea of "instantaneous rate of change" is the central idea of calculus. And of course it was invented by Newton and Leibnitz for just this purpose, to properly describe the motion of objects.

For simple cases, where the speed of the object is uniform and regular, we can use these relations to tell us things about how the object moved. For instance:

1. I have 1 hour to travel 30 km. How fast, in meters per second, must I travel?

$$\Delta s = 30 \text{ km} * \left(\frac{1000 \text{ m}}{1 \text{ km}} \right) = 30,000 \text{ m}$$

$$\Delta t = 1 \text{ hour} * \left(\frac{60 \text{ min}}{1 \text{ hour}} \right) * \left(\frac{60 \text{ s}}{1 \text{ min}} \right) = 3600 \text{ s}$$

$$v_{\text{average}} = \frac{\Delta s}{\Delta t} = \frac{30,000 \text{ m}}{3600 \text{ s}} = 8.3 \frac{\text{m}}{\text{s}}$$

2. I can jog at about 2.2 m/s, and I continue for 3 hours, how far do I get?

$$\Delta t = 3 \text{ hours} * \left(\frac{3600 \text{ s}}{1 \text{ hour}} \right) = 10,800 \text{ s}$$

$$v_{\text{average}} = \frac{\Delta s}{\Delta t} \quad \text{so}$$

$$\Delta s = v_{\text{average}} \Delta t = 2.2 \frac{\text{m}}{\text{s}} * 10,800 \text{ s} = 23,760 \text{ m} = 23.8 \text{ km}$$

How fast is fast? We will usually describe speeds in this class in units of meters per second. It is useful to get a mental image of speed by comparing to more familiar units. Most of us are familiar with the mile-per-hour scale for speed because we use cars all the time. By watching the speedometer and looking at motion, you get some visceral sense of velocity. As a result, it is often helpful to note that:

$$1 \frac{\text{m}}{\text{s}} = 2.24 \text{ mph} \quad \text{so} \quad 45 \frac{\text{m}}{\text{s}} \approx 100 \text{ mph}$$

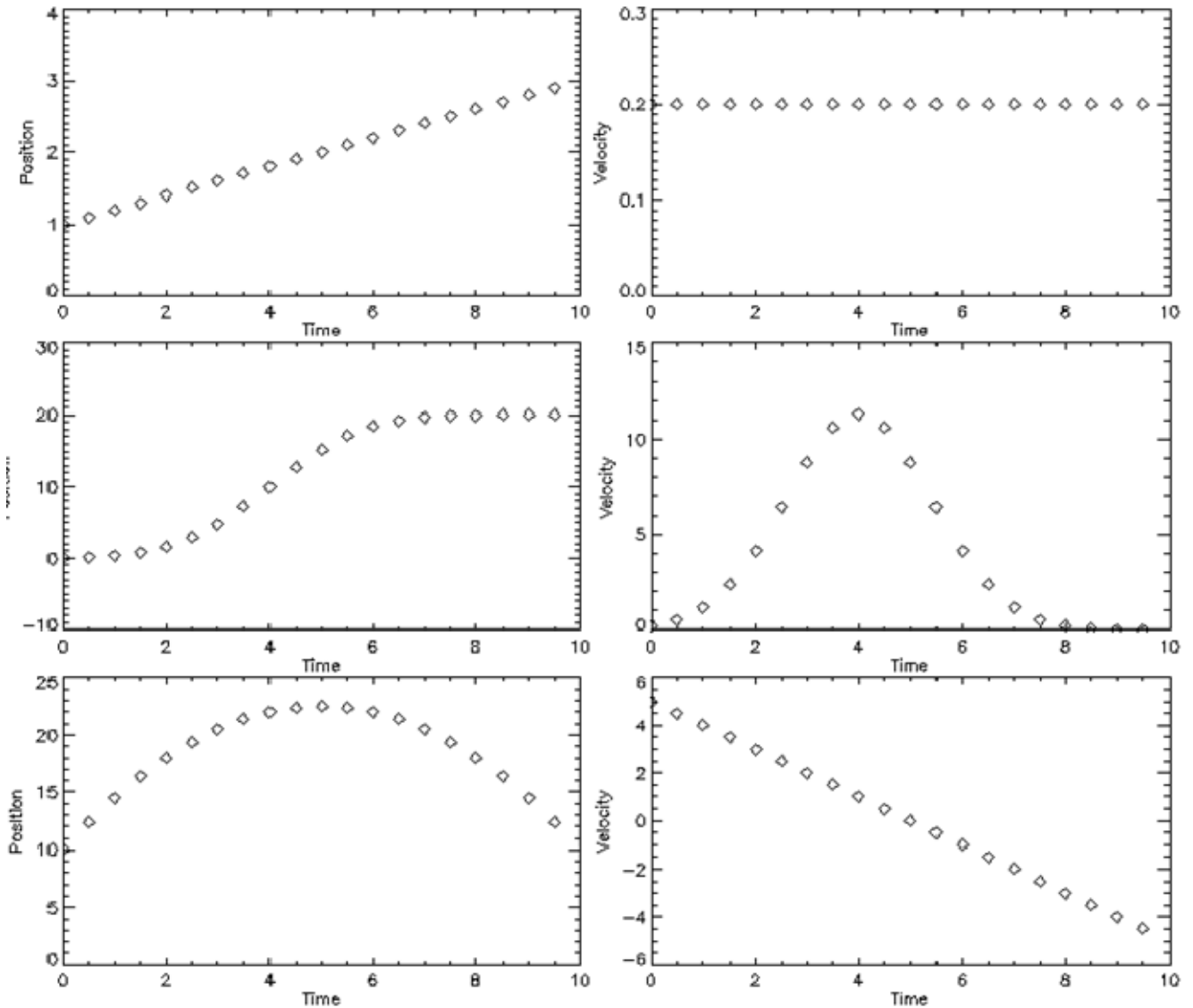
Here are some example speeds, just to give you a sense of the range we might deal with:

- Sea floor spreading: 1×10^{-9} m/s (~ 3 cm/yr)
- Grass growing: 2×10^{-8} m/s
- Glacier: 3×10^{-6} m/s
- Walking: 1.3 m/s
- Car: 25 m/s
- Sound in air: 330 m/s
- Earth's motion around the sun: 2.9×10^4 m/s
- Sun's motion around the Milky Way center: 2.2×10^5 m/s
- Approximate speed of an electron orbiting in Hydrogen: 2×10^6 m/s
- Speed of light in empty space: 2.998×10^8 m/s (~ 1 ft/ns)

Speed-Time Graphs and finding them from position time graphs:

If there are variations in speed, then to accurately describe the motion we have to consider the instantaneous speed. Using the instantaneous speed means we have a measure of the speed at each instant. So now our description of motion includes both a position s_i and a speed v_i at each instant t_i . This means we should be able to make a speed time graph, just as we have a position time graph.

What is the relation between the two? The instantaneous speed ds/dt also defines the slope of the position-time curve, so to create a speed time graph from a position time graph, you have to examine the slope of the position time graph at each point in time, and put that slope on the velocity time graph. The figure below shows some simple examples of position time graphs (on the left) and their corresponding speed time graphs (on the right).

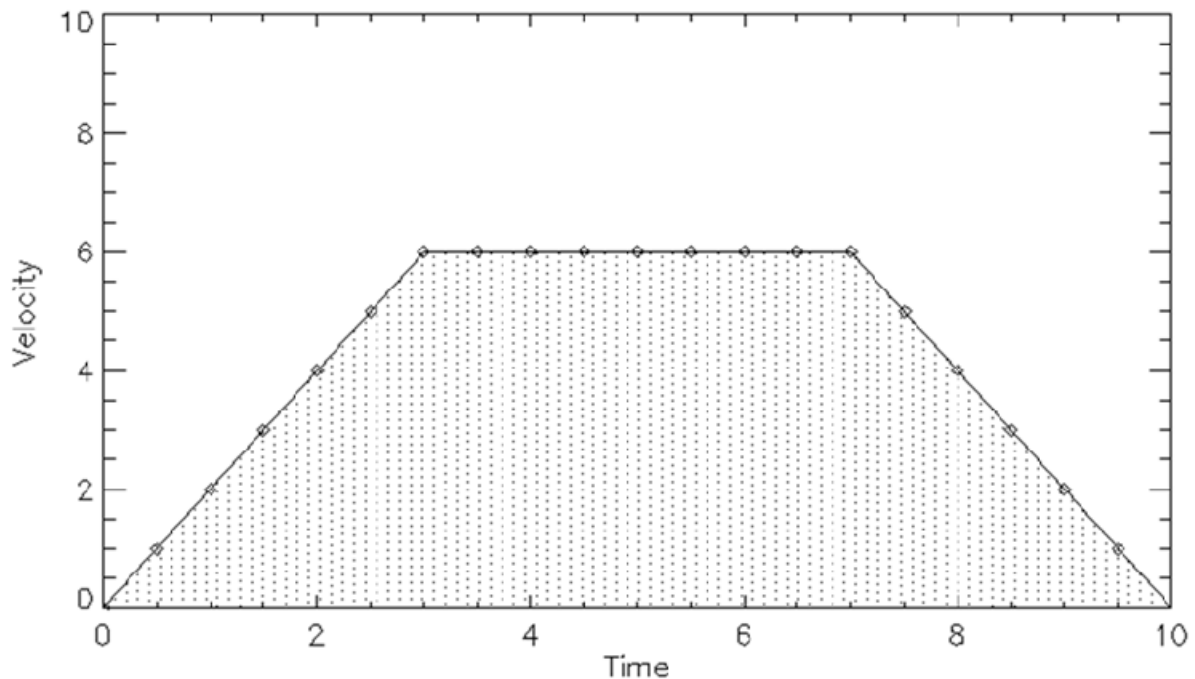


The slope of the position time graph determines the velocity time graph. In particular, any "linear" position time graph corresponds to a constant velocity.

Finding distance from velocity-time graphs:

Is there a way to reverse this? Given the velocity time graph, can I determine the position time graph? If I know how fast you are going at each instant, I can determine how far you went in each little period of time. Take the velocity in each second, multiply by 1 second to get how far you went, and add this to the total.

What does this mean on the velocity-time graph?



As I take the velocity in each period, and multiply by the time between samples, what I get is the area under the velocity time curve. So the distance traveled is just the area under the velocity time curve. In this case this area is:

$$0-3 \text{ seconds} : 0.5 * \left(6 \frac{\text{m}}{\text{s}}\right) * 2 \text{ s} = 6 \text{ m}$$

$$3-7 \text{ seconds} : 6 \frac{\text{m}}{\text{s}} * 4 \text{ s} = 24 \text{ m}$$

$$7-10 \text{ seconds} : 0.5 * 6 \frac{\text{m}}{\text{s}} * 2 \text{ s} = 6 \text{ m}$$

$$\text{Total distance} = 6 \text{ m} + 24 \text{ m} + 6 \text{ m} = 36 \text{ m}$$

This "area under the velocity time graph" is just an expression of the integral of the velocity time curve:

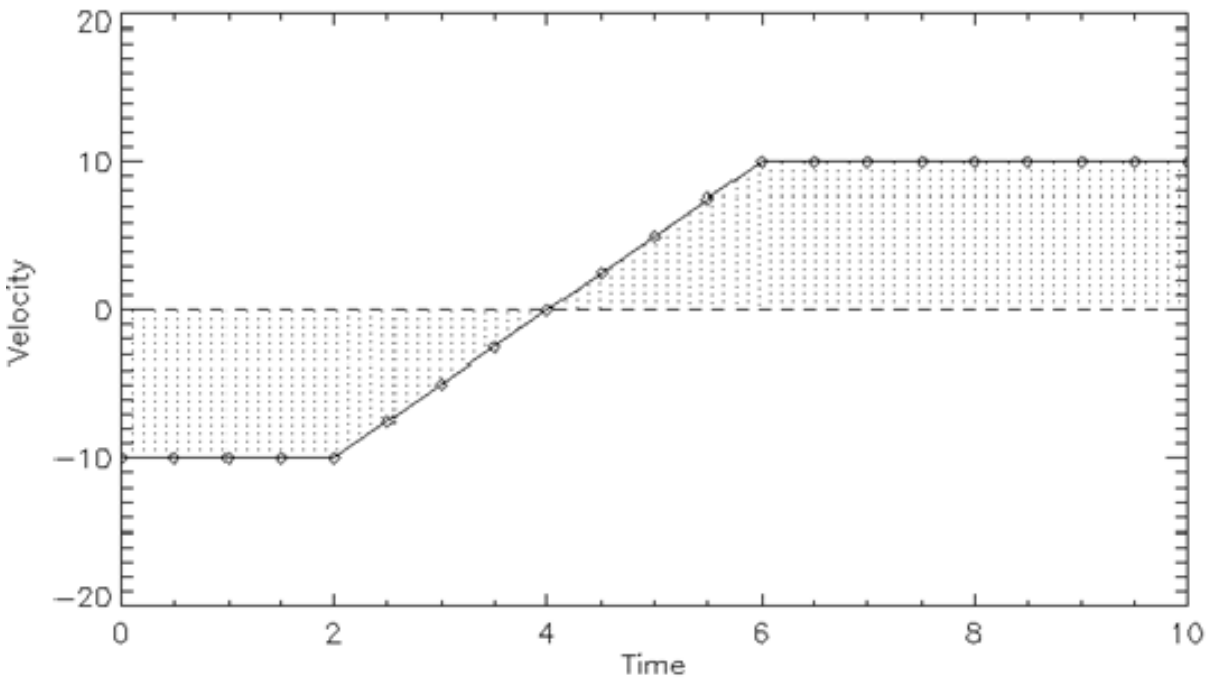
$$\text{Area} = \int_0^{10} v_{\text{instantaneous}}(t) dt = \int_0^{10} \frac{ds(t)}{dt} dt$$

So, we have a pair of relations:

$$s(t_i) - s(t_0) = \int_{t_0}^{t_i} v(t) dt$$

$$v(t_i) = \left. \frac{ds(t)}{dt} \right|_{t_i}$$

What if the speed is negative?



In this case, the part with negative velocity means motion in the negative direction, so it adds in "negative area" to the total:

$$0 - 2 \text{ seconds: } -10 \text{ m/s} \times 2 \text{ s} = -20 \text{ m}$$

$$2 - 4 \text{ seconds: } 0.5 \times -10 \text{ m/s} \times 2 \text{ s} = -10 \text{ m}$$

$$4 - 6 \text{ seconds: } 0.5 \times 10 \text{ m/s} \times 2 \text{ s} = 10 \text{ m}$$

$$6 - 10 \text{ seconds: } 10 \text{ m/s} \times 4 \text{ s} = 40 \text{ m}$$

$$\text{Total distance} = -20 \text{ m} + -10 \text{ m} + 10 \text{ m} + 40 \text{ m} = 20 \text{ m}$$

Negative area; what does that mean? Well it isn't really negative area, it's just distance traveled to the left instead of to the right.

Now that we understand the relations between position and velocity formally, we can determine precisely the velocity vs. time from a description of position vs. time. Imagine that we can describe the position as a

function of time as some function $s(t)$. We find the velocity vs. time by taking the derivative of this function:

If $s(t) = 15 + 0.5t + 0.1t^2$ then

$$v(t) = \frac{ds(t)}{dt} = 0.5 + 0.2t$$

Working in the opposite direction:

If $v(t) = 0.5 + 0.2t$ how far does this travel from $t = 0$ to $t = 4$?

$$s(t) = \int_0^4 v(t) dt = (0.5t + 0.1t^2) \Big|_0^4 = (0.5 * 4 + 0.1 * 4^2) - (0.5 * 0 + 0.1 * 0^2) = 3.6 \text{ m}$$

6.3 Changes in velocity and acceleration

We have talked about ways to relate the position, velocity, and time, for motion of objects. Earlier in the course, Newton's first law taught us that there is no real need to explain motion, but it is necessary to explain *changes* in motion. To describe these changes in motion we will need to include a way to describe changes in velocity.

We described the rate of change in position with time using velocity:

$$v = \lim_{\Delta t \rightarrow 0} \frac{\Delta s}{\Delta t} = \frac{ds}{dt}$$

And we describe the rate of change in velocity with time as acceleration:

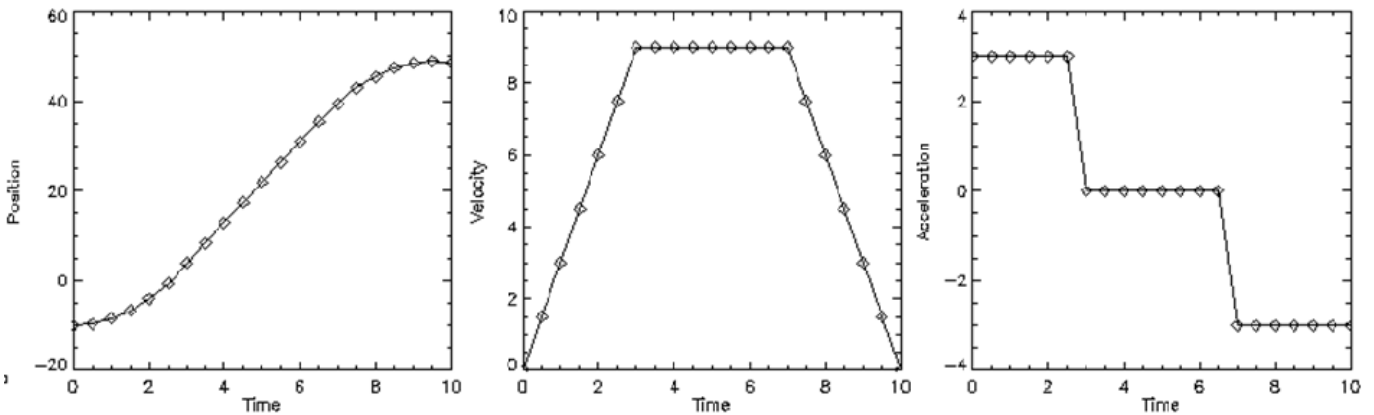
$$a = \lim_{\Delta t \rightarrow 0} \frac{\Delta v}{\Delta t} = \frac{dv}{dt}$$

What are the dimensions of this? It's a change in velocity (which has units of distance/time) divided by a time, so acceleration has dimensions of (distance/time) / time = distance / time². In the usual units acceleration is expressed in meters per second².

Acceleration vs. time histories

We begin our consideration of one dimensional motion by thinking about a sprinter moving in a straight line. For this discussion $\Delta s > 0$ means displacement to the right $\Delta s < 0$ means displacement to the left.

The sprinter begins at rest. She speeds up until she is moving at a constant speed, which she continues for a while. Then after crossing the finish, she gradually slows down and stops. How do we describe her motion? Think about this motion in terms of your car. You start on a straight road, speed up to some speed, travel at a constant speed for a while, then slow to a stop. What does your speedometer read during this time? What in your car measures the distance? Is there anything in your car which measures acceleration?



We now have a three-fold history of the motion of the object. Note that although the velocity here is always positive, the acceleration is both positive and negative at different times. This is an essential point: that although velocity is always along the direction of motion, acceleration can be in any direction.

Notice one more thing as well. In this example, the acceleration basically takes on three constant values; one positive, one zero, and one negative. What is the acceleration in between these three values? In the idealized model we have here, the acceleration changes from positive to zero, and from zero to negative, instantly. But of course in a real, physical case this would not be the case. Instead, the transitions from one constant value of acceleration to the next would be smooth and continuous, rather than infinitely sudden and discontinuous.

Finding change in velocity from the acceleration vs. time history

Just as the change in position Δs can be found from the area under the velocity vs. time history, the change in speed Δv can be found from the area under the acceleration vs. time history. When the acceleration is positive, the velocity becomes more positive during each interval of time. When the acceleration is negative, the velocity becomes more negative during each interval of time.

We don't say that the speed increases when the acceleration is positive because this may not be true. Imagine an object traveling very fast in the negative direction. A positive acceleration implies that this velocity becomes more positive in each time interval, but since it begins large and negative, the speed becomes smaller as time goes on. Once again, position always changes in the direction of the velocity, and velocity always changes in the direction of acceleration.

What can we say about the relations between distance, velocity, and acceleration when the acceleration is not a constant? There are essential general relations that always apply. What we *always* know is how the position-time, velocity-time, and acceleration-time graphs are related to one another. There are four parts to remember:

1. The velocity is the slope of the position-time graph
2. The acceleration is the slope of the velocity-time graph
3. The distance traveled is the area under the velocity-time graph

4. The change in velocity is the area under the acceleration-time graph

These are true for every motion. Once you know any one of these three graphs, you can determine the other two. They are all very closely related.

Changes in acceleration; jerk, snap, and beyond...

The relations among position, velocity, and acceleration form an obvious pattern:

$$\begin{aligned} s(t) \\ v(t) &= \frac{ds(t)}{dt} \\ a(t) &= \frac{dv(t)}{dt} = \frac{d^2s(t)}{dt^2} \end{aligned}$$

Why not continue this pattern? If velocity is the time rate of change of the position, and acceleration is the time rate of change of velocity, what is the time rate of change of acceleration? And what is the time rate of change of the time rate of change of the acceleration? We could talk about these things, and in fact they have names. The time rate of change of the acceleration is called 'jerk', and the time rate of change of jerk is called 'snap' or sometimes 'jounce'...

$$\begin{aligned} \text{jerk}(t) &= \frac{da(t)}{dt} = \frac{d^2v(t)}{dt^2} = \frac{d^3s(t)}{dt^3} \\ \text{snap}(t) &= \frac{d(\text{jerk}(t))}{dt} = \frac{d^2a(t)}{dt^2} = \frac{d^3v(t)}{dt^3} = \frac{d^4s(t)}{dt^4} \end{aligned}$$

Obviously this could continue forever. Why do we usually stop with acceleration? We needn't, but this proves reasonable to do, mostly for two reasons.

First, to predict the full path of an object, we need to know how it is moving and how that motion changes. We have seen that if we know the velocity at all times, we can predict the position, and that to know the velocity at all times, we need only know the acceleration at all times. So if we actually know the acceleration at all times, we can predict the full motion of objects. It's true that the acceleration may change with time (so that there would be jerk, and perhaps snap and beyond), but all we need to know to predict the path is the acceleration.

Second, we have found that there is a profound connection between acceleration and the forces which cause changes in motion. Remember Newton's second law:

$$\vec{F}_{total} = \frac{d\vec{p}}{dt}$$

The momentum $\vec{p} = m\vec{v}$, and since in most cases the mass of an object does not change,

$$\vec{F}_{total} = \frac{d(m\vec{v})}{dt} = m \frac{d\vec{v}}{dt} = m\vec{a}$$

Seen in this context, Newton's second tells us that forces cause accelerations. If we know all the forces, we can predict the accelerations, and as we have just argued, knowing the accelerations allows us to fully predict the motions of objects.

So while jerk, snap, and even higher derivatives of the motion of an object exist, they aren't the source of change in the world. Forces cause the changes we see, and forces, most basically, create accelerations. Once we know the accelerations, we can predict the motion of everything. So position, velocity, and acceleration will be the focus of our study of motion.

6.4 Modeling some simple motions: constant position, constant velocity, and constant acceleration

Most motion is complex, but as usual we will begin by analyzing in detail some very simple cases. The first, very common, is also very trivial; constant position. In this case:

$$s(t) = s_0 \quad v(t) = 0 \quad a(t) = 0$$

Most of the things you see around you right now appear to be exhibiting this kind of motion. But are they really? After all, the Earth spins on its axis once a day, and orbits the Sun once a year. Meanwhile the Sun and the rest of the solar system orbit the center of the Milky Way about once every 230 million years. So these things, and you, are not *really* at rest. They are moving, and their motions change with time, though these changes happen rather slowly. Since the changes in motion associated with the spinning and orbital motion of the Earth happen slowly, we can approximate them by motion with constant velocity. How would we describe position, speed, and acceleration for constant velocity motion?

$$s(t) = s_0 + v_0 t \quad v(t) = v_0 \quad a(t) = 0$$

Notice that the constant position motion we considered first is really just a special case of constant velocity motion; the one with $v_0 = 0$. While you sit in your room reading this, everything you see is moving with a nearly constant velocity. You don't notice this motion because you are moving in the same way; the relative velocity between you and your surroundings is zero, and it's this zero relative velocity which remains unchanged with time.

Is motion with constant velocity a common thing? When something moves with constant velocity it just keeps going, never staying near anything which is not moving along with it. There are many cases of motion with constant velocity which last *for a while*. We have already seen a good example; terminal velocity, in which one force pushing something forward is balanced by another force which resists the motion. Motions of living things are often like this, at least for a while: people walking down the street, birds flying through the air, swimming fish, all have periods in their motion during which a balance of forces leads to constant velocity. So motion with a constant velocity will be a model we will use to describe the motions of some objects *for a while*.

The next simplest example we might consider is motion with constant acceleration. This is often described in introductory physics courses as a common situation, though in fact it is not. What would happen to something which experienced a constant acceleration? Pretty quickly, it's going really fast, and then of course it doesn't stay around long. To have constant acceleration, the total force on an object must be constant, somehow continuing to act on the object while it goes zooming off through space, moving faster and faster. When you think about it this way, it's apparent that this is not so easy to arrange. So in fact, motion with constant acceleration is exceptionally rare.

Why talk about it then? There are two reasons really. First, it sometimes does happen *for a little while*, at least approximately. So we can use the equations which describe motion with constant acceleration as a model for more complex motions, at least during the short periods when the acceleration is approximately constant. The second reason for talking about motion with constant acceleration is because this kind of motion is easy to solve analytically. The latter is probably the main reason the example is so popular in introductory physics courses.

Remember that we have already considered one special case of motion with constant acceleration: motion in which the acceleration is zero, so the velocity never changes. Now we want to consider cases where the acceleration is not zero, but still constant.

$$a = \frac{dv}{dt} = a_0$$

In this case, with constant acceleration, there is no difference between average acceleration and instantaneous acceleration. We can write:

$$a_0 = \frac{dv}{dt} = \frac{\Delta v}{\Delta t} \quad \text{so} \quad \Delta v = a_0 \Delta t$$

or:

$$\Delta v = v_f - v_i = a_0 \Delta t \quad \text{or} \quad v_f = v_i + a_0 \Delta t$$

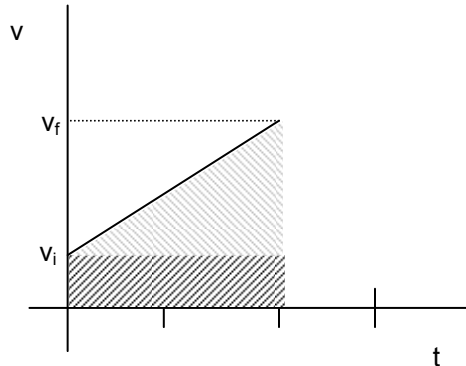
In this simple case we can determine how the velocity of an object changes completely. Taking a sample case in which the initial velocity is 5 m/s, and the acceleration is 3 m/s². What is the velocity after 10 seconds?

$$v_f = v_i + a_0 \Delta t = 5 \frac{\text{m}}{\text{s}} + 3 \frac{\text{m}}{\text{s}^2} \times 10 \text{ s} = 35 \frac{\text{m}}{\text{s}}$$

We might instead ask: if a car can accelerate at 5 m/s², how long does it take it to accelerate from a stop to 35 m/s?

$$v_f = v_i + a_0 \Delta t \quad \text{or} \quad \Delta t = \frac{v_f - v_i}{a_0} = \frac{35 \frac{\text{m}}{\text{s}} - 0 \frac{\text{m}}{\text{s}}}{3 \frac{\text{m}}{\text{s}^2}} = 7 \text{ s}$$

So now we know how the velocity changes. What about distance traveled? If I start at some velocity v_i , and accelerate at a constant rate to some velocity v_f for some total period of time Δt , I can draw the velocity-time curve for this as:



Now remember that because for each little period dt , the distance traveled is $ds = v(t)dt$, the total distance traveled is equal to the area under the velocity-time curve. This area can be determined by breaking the above into two parts: a lower rectangle which has area:

$$\text{area} = v_i \Delta t$$

and an upper triangle which has an area:

$$\text{area} = \frac{1}{2}(v_f - v_i) \Delta t$$

So the total distance traveled is :

$$\Delta s = v_i \Delta t + \frac{1}{2}(v_f - v_i) \Delta t = v_i \Delta t + v_{av} \Delta t$$

This should not be too surprising. The first term in this equation is just the distance the object would have traveled if the velocity were not changing. The second term represents the additional distance traveled because the velocity is increasing.

It is often useful to rearrange these relations, expressing them in different ways. The relation we just wrote is very useful if you know the initial and final velocities, and you want to know how far you have traveled. If instead you know your initial velocity and your acceleration, and you want to know how far you go, it is more useful to restate the above relation in terms of acceleration:

$$a = \frac{(v_f - v_i)}{\Delta t} \quad \text{or} \quad (v_f - v_i) = a \Delta t$$

this lets us replace this part of the above relation to find:

$$\Delta s = v_i \Delta t + \frac{1}{2} a (\Delta t)^2$$

Remembering that $\Delta s = s_f - s_i$, we can write this in a commonly used and explicit form:

$$s_f = s_i + v_i \Delta t + \frac{1}{2} a (\Delta t)^2$$

Equations for constant acceleration motion derived using calculus

Let's determine these relations more directly, using some simple calculus

$$a = \frac{dv}{dt}$$

so that

$$dv = a dt \quad \text{or} \quad \int dv = \int a dt = a \int dt$$

We were able to take the acceleration out of the integral because we have assumed it is constant. This form is easily solved to find:

$$v_f = a(t_f - t_i) + C$$

where C is a constant of integration, something undetermined by the mathematics. In physics, such constants must always be determined by the physical circumstances, by the "boundary conditions" of the problem. In this case, we have to know what the velocity was at time $t = t_i$. If this is written v_i then we know that:

$$v_i = a(t_i - t_i) + C$$

so:

$$C = v_i$$

And we write:

$$v_f = a(t_f - t_i) + v_i$$

In a similar way we write:

$$v = \frac{ds}{dt} \quad \text{or} \quad ds = v dt$$

So

$$\int ds = s = \int v dt = \int (at + v_i) dt = \frac{1}{2} a (t_f - t_i)^2 + v_i (t_f - t_i) + C$$

where C is a new constant of integration. Just as above, we can determine this by applying the boundary condition that $s = s_i$ when $t_f = t_i$, we have:

$$s_f = s_i + v_i (t_f - t_i) + \frac{1}{2} a (t_f - t_i)^2$$

Now we have two kinematic equations to work with in situations where the acceleration is constant:

$$\begin{aligned} v_f &= v_i + a(t_f - t_i) \\ s_f &= s_i + v_i(t_f - t_i) + \frac{1}{2} a(t_f - t_i)^2 \end{aligned}$$

It is possible to combine these two to eliminate $(t_f - t_i)$ and set up a third kinematic equation. The derivation of this third equation is straightforward:

$$\begin{aligned} (t_f - t_i) &= \frac{(v_f - v_i)}{a} \\ (s_f - s_i) &= v_i \frac{(v_f - v_i)}{a} + \frac{1}{2} a \frac{(v_f - v_i)^2}{a^2} \end{aligned}$$

Or

$$2a(s_f - s_i) = (v_f - v_i)[2v_i + (v_f - v_i)] = (v_f - v_i)(v_f + v_i) = v_f^2 - v_i^2$$

to finally give:

$$v_f^2 - v_i^2 = 2a(s_f - s_i)$$

For convenience, it may be useful to lay out all three of these ‘kinematic equations’ for motion under the influence of constant acceleration together:

$$\begin{aligned} v_f &= v_i + a(t_f - t_i) \\ s_f &= s_i + v_i(t_f - t_i) + \frac{1}{2} a(t_f - t_i)^2 \\ v_f^2 - v_i^2 &= 2a(s_f - s_i) \end{aligned}$$

A caution

Very often in physics we follow a line of reasoning which begins with some simple assumptions and continue with it so far that we forget the assumptions we started with. What have we assumed about the acceleration to derive this set of equations? We have assumed that the acceleration is constant throughout!

These equations do not apply at all if the acceleration is not a constant. Why do we talk about the constant acceleration case? Mostly because it is easy to analytically solve, but also because it does sometimes happen, at least for a while. As a result, these equations can provide a useful element in a model meant to describe something more complex. The equations of motion for constant acceleration are another of our “spherical cow approximations”.

A Quick Summary of Some Important Relations

Completely general relations between position, velocity, and acceleration:

$$\begin{array}{ll} s(t) & \Delta s = \int_{t_i}^{t_f} v(t) dt \\ v(t) = \frac{ds(t)}{dt} & \Delta v = \int_{t_i}^{t_f} a(t) dt \\ a(t) = \frac{dv(t)}{dt} = \frac{d^2s(t)}{dt^2} & a(t) \end{array}$$

Put into words; the velocity at each instant is the slope of the position-time graph, and the acceleration at each instant is the slope of the velocity-time graph. The change in position during from an initial time t_i to a final time t_f is the area under the velocity-time graph between those two times. Likewise, the change in velocity during this time interval is the area under the acceleration-time graph between those two times.

Specific relations which predict position and speed when acceleration is constant:

$$\begin{array}{ll} v_f = v_i + a(t_f - t_i) & v_f = v_i + a\Delta t \\ s_f = s_i + v_i(t_f - t_i) + \frac{1}{2}a(t_f - t_i)^2 & s_f = s_i + v_i\Delta t + \frac{1}{2}a\Delta t^2 \\ v_f^2 - v_i^2 = 2a(s_f - s_i) & v_f^2 - v_i^2 = 2a\Delta s \end{array}$$

Here you see the same relations written in two different forms. In the first the initial position, time, and velocity are explicitly listed. In the second the change in time between t_f and t_i is written with the shorthand Δt , and the change in position is written with the shorthand Δs . You can of course use either form; just be sure you know that these apply *only* when acceleration is constant!

7. Getting started and moving around: what makes motion change

- 1) Quantifying a motion and its changes
 - i. Force and the alteration of motion
 - ii. Mass and momentum
 - iii. Momentum is relative, but changes in momentum are absolute
- 2) Force and the rate of change of momentum
 - i. Duration of force and impulse
 - ii. A simple example, the bouncing ball
 - iii. Force and acceleration
- 3) Weight and free fall
 - i. The idealized case with no other forces
 - ii. When is this idealization appropriate?
 - iii. Falling through air: speeding up to terminal velocity
- 4) Summing up one dimensional motion: getting started, traveling along, and stopping

Physics for the Life Sciences: Chapter #7

7.1 Force and the alteration of motion:

In the last lecture we learned how to describe motion with position time graphs, rate of motion with speed-time graphs, and changes in rate of motion with acceleration-time graphs. Now that we have the tools in place to describe motion we're going to learn about what causes motion to change, and how rapidly that change occurs.

Recall what Newton said, an object in motion remains in motion "unless compelled to change its motion by forces impressed upon it". The law of inertia tells us that every unmolested object remains in uniform motion. But right here Newton is telling us how motion changes; motion changes from uniform because of unbalanced forces.

Mass and momentum:

OK, forces alter motion. How do they do this quantitatively? We know from experience that the way in which a force acts on an object depends on its nature; it is easier to stop a running 2 year old than a hard charging 300 lb lineman. The proper way to express this inertia of motion, the ease or difficulty with which something is stopped, is through momentum, defined as:

$$\vec{p} = m\vec{v}$$

This is the mass of an object times its velocity. There are several things to notice about momentum:

- Momentum is a vector, so you need to be aware of both its magnitude and its direction. Changes in momentum will be vector changes. They can be changes in magnitude, or direction, or both.
- Momentum depends linearly on the mass of the object. Double the mass of a moving object and you double its momentum. Reduce the mass to a 10th of its original value and you reduce the momentum by a factor of ten.
- Momentum depends linearly on the magnitude of the velocity; double the velocity of an object and you double its momentum.

- The direction of momentum is always in the direction of the velocity (and hence always along the path of the object's motion).

This measure of how difficult it is to change the motion of an object is quantified in Newton's second law:

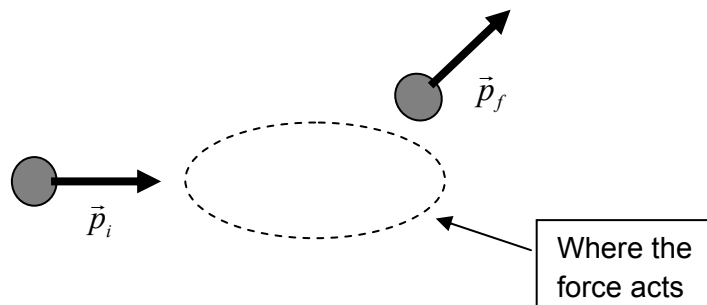
The force exerted on a body equals the resulting change in its momentum divided by the time elapsed in the process.

Expressed as an equation, this amounts to a quantitative definition of the force vector:

$$\vec{F}_{total} = \frac{d\vec{p}}{dt}$$

The total force on an object is equal to the change in its momentum divided by the time it takes to make that change. The dimensions of force are $[M(L/T)]/T = ML/T^2$ or in the usual units kgm/s^2 . This unit, 1 kgm/s^2 , is called a Newton for the obvious reason. How large is 1 N of force? It's about equal to the weight of a modest sized apple, like the one which fell on Newton's head in the (probably apocryphal) story.

What is a force really, what is its essential nature? A force is a fundamental thing, the most basic description of an interaction between two objects. Because it is so fundamental, we can't really describe what it is in terms of other things. Our best way to describe forces is to talk about what they do. In fact the study of forces and how they behave is a very large fraction of physics, and in the end, they *are* what they do. What do I mean? We can tell there is a force present only by seeing it act on an object. This is illustrated in the figure below. When we see the momentum of an object change, we know that a force has acted. In a real sense the change in momentum and the force are the same thing; they are equal.



We don't have to know in any detail what happens in the oval to describe the force; we only have to know how the momentum of the object changes. When we see the momentum change, we know a force acted.

Relativity and momentum:

There is an important point to note here. Because momentum depends on velocity, its value is not absolute; it will be different for different observers. An example should help to illuminate this. Imagine a thoughtless motorist, speeding down a country road at 45 miles per hour. He holds a can of Dr. Pepper,

takes the last swig, then gently flips the empty can out the open window. A peaceful pedestrian, minding his own business, is promptly struck in the forehead by a can traveling at 45 mph.

To the driver, the can moves slowly, and has a small momentum directed straight out the car window. To the poor pedestrian walking beside the road, the can is moving very fast, and has a large momentum, mostly in the direction of motion of the car. Which observer is right? How might the can have both large and small momentum?

In fact both observers are correct, because the absolute value of an object's momentum has no meaning for physics. What physics cares about, as Newton's second law tells us, are *changes* in momentum. There's no problem with two observers seeing the can moving in different ways. What they must agree on is what happens when forces act; they must agree about how the momentum of the can changes. What do the two observers see when the can strikes the innocent forehead of our pedestrian?

The driver sees the can begin with a small momentum to the side, then suddenly acquire a large momentum in the backward direction. The momentum changes a lot. This change is given by:

$$\Delta\vec{p} = \vec{p}_f - \vec{p}_i$$

So the driver sees a small initial momentum and a large negative final momentum. So the momentum change is large and in the direction opposite the car's motion. The pedestrian sees the can come toward him with a large positive initial momentum. Then it strikes his forehead and comes to a stop, with a nearly zero final momentum. So the pedestrian also sees a change in momentum which is large and in the direction opposite the car's motion.

The value of the momentum of an object is relative; it can truly be different when seen by different observers. Changes in momentum, however, are absolute, and will be seen as the same by all observers. Note that this is related to the law of inertia; objects in motion continue in motion; the only relevant thing is *changes* in motion, not the absolute amount of motion something has. It is **change** that is important in physics.

There are times when two observers see different changes in momentum, but this happens *only* when one or both of the observers are themselves changing their motion. Imagine, for example, that you place your coffee cup on the dashboard of a car which suddenly accelerates forward (carrying you with it). It will seem to you that the cup is thrown backward in the car; that its momentum suddenly increases in a negative direction. An observer standing on the sidewalk sees things differently. She sees the cup remain in place, while you and the car suddenly accelerate forward. You and the person on the sidewalk disagree about how the momentum of the cup changed, but this can only occur if one (or both) of the observers are themselves changing their motion.

When observers travel at constant velocity, when their motions do not change, they will *all* observe the same changes in momentum for all objects. They will agree on what forces act, and be able to explain them in coherent ways. These observers are called 'inertial' observers. If one or both of the observers is changing their motion, they may not agree on the changes in momentum they see. These observers would think that different forces were acting, and could not really agree about what is happening. Such observers

are called ‘non-inertial’. In general, we will analyze what happens from the perspective of inertial observers. But we should always remember that our own motions are ever changing; the Earth rotates every day, and orbits the sun every year. So although we are *nearly* inertial observers, in detail we are not.

7.2 Duration of force and impulse:

There is an alternate way of looking at Newton’s second law which is quite instructive. If we rearrange this equation as:

$$\vec{F}\Delta t = \Delta\vec{p}$$

we can explicitly see that to obtain a certain change in momentum we can either use a large force for a short time, or a small force for a long time. The product of force times time is what creates momentum change. This product is called “impulse”, and in some ways its action is easier to understand than force.

We can turn this around productively. If we see a certain momentum change, we know how large the impulse was. Then, if we know about how long the interaction was happening, we can estimate the average size of the force.

Here’s a simple example. I run to the East at 2.5 m/s. Then, rather quickly, in a half second or so, I stop, turn around, and run back to the West at 2.5 m/s. About how large is the force which must act on me to alter my motion in this way? Newton’s second law gives us what we need to find out. How does my momentum change?

$$\Delta\vec{p} = \vec{p}_f - \vec{p}_i$$

My initial velocity is 2.5 m/s East. Let’s call East the positive direction, and West negative. My mass is about 80 kg. So my initial momentum is:

$$\vec{p}_i = 80 \text{ kg} \times 2.5 \text{ m/s} = 200 \text{ kgm/s } \hat{E}$$

The final momentum has the same magnitude, but is in the opposite direction:

$$\vec{p}_f = 80 \text{ kg} \times -2.5 \text{ m/s} = -200 \text{ kgm/s } \hat{E}$$

Putting these together, we get

$$\Delta\vec{p} = \vec{p}_f - \vec{p}_i = -200 \text{ kgm/s } \hat{E} - 200 \text{ kgm/s } \hat{E} = -400 \text{ kgm/s } \hat{E}$$

In this answer, the minus sign simply means in the negative, or West, direction. Now we get the force:

$$\vec{F} = \frac{\Delta\vec{p}}{\Delta t} = \frac{-400 \text{ kgm/s } \hat{E}}{0.5 \text{ s}} = -800 \text{ kgm/s}^2 \hat{E} = -800 \text{ N } \hat{E}$$

This just means 800 N in the West direction. It's not surprising that to change my motion from going East to going West requires a force in the West direction. This is a pretty large force too. Remembering that 1 N is about the weight of an apple, this is the weight of 800 apples. It's also about equal to my weight, which is $80 \text{ kg} \times 9.8 \text{ m/s}^2 = 784 \text{ N}$.

If I'm going to run forward at 2.5 m/s (6.25 mph, a modest jog) and turn around in half a second, there has to be a force equal to my weight acting in the opposite direction for about a half a second. Where does this force come from? I push my feet against the ground, and the ground pushes back on me. It's the force of the ground pushing on me that turns me around and sends me back the other way. Of course that frictional force only appears because I push on the ground.



Force and acceleration:

This view that force is coupled to changes in momentum is the fundamental way of stating Newtonian mechanics, and it is the most correct. But there is another, often useful, way of looking at force.

$$\vec{F} = \frac{d\vec{p}}{dt} = \frac{d(m\vec{v})}{dt} = m \frac{d\vec{v}}{dt} = m\vec{a}$$

Now to derive this we have had to assert that the mass doesn't change with time, so that

$d(m\vec{v}) = md(\vec{v})$. Is this true? For a very long time, everyone thought so, until early in this century when the theory of relativity, and more important its experimental confirmation, demonstrated that it was not. When objects approach the speed of light, the **only** correct formulation is

$$\vec{F} = \frac{d\vec{p}}{dt}$$

And it is impossible to say that the mass does not change. We should make it completely clear that $\vec{F} = m\vec{a}$ is perfectly acceptable for everything we will do in this book, but it is important for you to understand why it is that we emphasize the time-rate-of-change of momentum form. This conception of the second law, which was Newton's original conception, is now known to be the only correct one for all speeds.

We will work a lot from now on with the more approximate $\vec{F} = m\vec{a}$ formulation. It's important, because it allows us to determine forces from observed accelerations, and accelerations from known forces. Doing analyses like this will be the subject of the next several sections of the text.

7.3 Weight: the force exerted by gravity

This alternate version of Newton's second law ($\vec{F} = m\vec{a}$) lets us see weight in a new light. In Chapter 3 we showed that the usual way of describing the downward pull of the Earth on objects near its surface actually stems from a more general theory of gravitational interaction which also governs the motion of planets.

$$W = F_{\text{Earth-object}} = \frac{GM_{\text{Earth}}m_{\text{object}}}{R_{\text{Earth}}^2} = m_{\text{object}} \left(\frac{GM_{\text{Earth}}}{R_{\text{Earth}}^2} \right) = m_{\text{object}} g$$

In this equation, the constant ‘g’ is determined from the gravitational strength constant G, the mass of the Earth, and the radius of the Earth. It has a value of 9.8 m/s².

Notice that g has the units of acceleration. If weight is the *only* force acting on an object, it will accelerate toward the center of the Earth at a rate of 9.8 m/s². Of course if other forces also act, the acceleration may be different from this. And in fact it is very rare that no other forces act. You, for example, sit in your chair. Your weight pulls you down, and in the absence of other forces, you would accelerate downward at 9.8 m/s². But of course other forces balance this downward force, and you remain in place.

Imagine you drop an object through the air. At the first instant, the moment you release it, the only force acting on it is its weight, and indeed at this moment it accelerates downward with an acceleration of 9.8 m/s². As soon as it begins to move, however, its downward motion is resisted by an upward frictional force. Once this begins, the *total* force on the object becomes less and less, its acceleration becomes smaller and smaller, until eventually it no longer accelerates and falls from then on with a constant, terminal velocity.

Falling objects and idealized free fall:

The fact that falling objects experience friction which resists their motion created confusion about the nature of weight and falling objects for thousands of years. The question of why things fall when dropped is as old as our ability to form it. It touches on our most basic understanding of how the universe works.

In ancient Greece, it was thought that everything was made of four elements: earth, air, fire, and water. These elements had distinctive properties. Earth and water were both imbued with gravity, while air and fire were possessed of levity. Most ancient Greeks believed that heavy objects (made of earth or water) fell downward simply because it was in their nature to do so. This falling motion was *intrinsic* to the object. Likewise fire and air might rise simply because it was in their nature to do so.

Our view today is very different. Newtonian theory suggests that an object falls because of an interaction between it and the Earth. All of Newtonian theory relies on this idea of interaction. Nothing ever happens on its own; everything is the result of some kind of interaction between things. Instead of the object falling on its own, it falls because the Earth makes it fall. It is not surprising that the ancients didn’t speak of this interaction. After all, gravity is a non-contact force. You can’t *see* any interaction taking place between the Earth and a ball you drop.

Another ancient view was the Aristotelian idea that objects fall with constant speeds proportional to their weights. This was not a crazy idea. In the presence of air friction it is true that many dropped objects quickly reach a "terminal" velocity which depends on their weight and shape. We see this as the result of another force (friction) acting in addition to weight, but again, this is not so obvious, especially for objects moving through the air.

Galileo was responsible for successful public refutation of the idea that heavier objects fall faster. His methods of showing this form a model for how we argue all points in science. He brought several important techniques to bear:

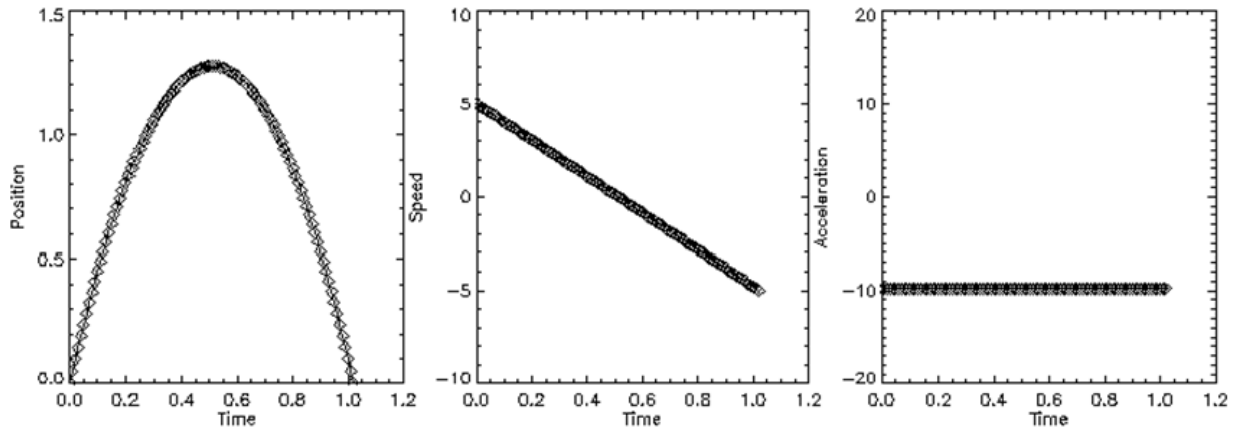
1. **Idealization:** Galileo could see that friction was confusing the problem, and studied ways to remove it. He could not remove air friction (vacuum pumps were not invented until just after he died), but he could reduce it by using smooth, heavy balls, and restricting their motion to relatively low velocities. The notion of focusing on simple, idealized, examples has proven powerfully useful in physics.
2. **Experiment:** Experiments are the ultimate arbiters of truth in science. It might seem a simple matter to test the Aristotelian model directly. But making measurements in the 17th century was very different from today. To appreciate this, you might ask how Galileo measured time. He had no quartz watches or atomic clocks. Instead he had to use pendulum clocks or design water clocks which would drip at regular rates. They weren't very precise, and this made it very difficult for him to measure the time it takes to drop a ball. To avoid this problem he had to devise ways to "dilute" the acceleration due to gravity. For this purpose, he replaced dropping a ball with rolling it down a very smooth, gradually sloping ramp. This slower motion gave him the ability to measure more precisely, and also reduced the influence of air friction on his results.
3. **Thought experiments:** Sometimes a real experiment is difficult to conduct, but one can imagine what it might be like. This is sometimes called doing a 'thought experiment', something which many leading scientists become very good at. Galileo was particularly brilliant in this way. In the Aristotelian model a heavy rock would fall faster than a light one. Galileo imagined tying the small, slowly falling stone to the heavy, rapidly falling one. On one hand, you might expect the smaller, slowly falling stone to slow down the heavy, rapidly falling one. On the other, you might expect the combination, which is larger than either of the two, to fall faster still. Both predictions seemed reasonable, and their disagreement was enough to convince Galileo that the Aristotelian principle must be wrong.

Once Galileo removed the effects of friction (in fact and in his mind), he found that all objects, independent of their mass, accelerated downward at the same rate. Gravity creates in any unsupported object an acceleration which is independent of its properties. This is an extraordinary fact, almost bizarre. How can it be that all objects, independent of their properties, behave the same way? This is egalitarianism indeed. It was a clue that the size of the downward force exerted on an object by the Earth depended on its mass. This meant that a small object would receive a small force, while a large object would receive a large force; each would get just enough to make it accelerate downward at precisely the same rate.

Free fall and motion under constant acceleration:

Galileo's basic observation was that in the absence of air friction all objects, independent of their properties, are accelerated towards the center of the Earth at the same rate. The magnitude of this acceleration is 9.8m/s^2 , and it is directed towards the center of the Earth. While this is never precisely true here on Earth, we can use this as a simple, first approximation model to discuss the motion of a body moving straight up or down quantitatively.

A simple example of this motion is when you toss a ball straight up into the air. This motion begins with some positive speed which is continually, regularly, reduced by the constant negative acceleration of gravity, until at the top of the path, the speed becomes zero. Then as it descends it has a continually, regularly, increasing negative speed. All the while the acceleration has a constant value of 9.8 m/s^2 downwards. Position-time, speed-time, and acceleration-time graphs for motion like this are shown below. In this case, the initial upward velocity is 5 m/s .



This ball, moving in the absence of any air friction, would rise about 1.3 m above your hand, then fall back down, all during a period of about one second. During this time, the speed of the ball would start at $+5 \text{ m/s}$, fall to zero, then increase to a negative speed of -5 m/s . Throughout the whole motion the acceleration of the ball would be a constant -9.8 m/s^2 , the acceleration due to gravity.

It is sometimes difficult to believe that there is still acceleration at the moment when the ball comes to a stop at the top of its path. It's perhaps worthwhile to consider the following thought experiment. Imagine you are driving your car up a hill and you slip it into neutral, coasting upward. You gradually slow to a stop. What happens if, at just the instant when your car stops, you abruptly put on the brake? What if, instead, you are coasting on a flat road and put on the brake at just the moment at which you stop? The fact that you experience a "jerk" when coasting up a hill is evidence that by putting on the brake you actually change your acceleration (from a constant negative value to zero) at that instant.

In the last chapter we discussed general motion under constant acceleration and developed a set of equations to describe it:

$$v_f = v_i + a\Delta t$$

$$s_f = s_i + v_i\Delta t + \frac{1}{2}a\Delta t^2$$

$$v_f^2 - v_i^2 = 2a(s_f - s_i)$$

To use these equations here, we need only remember that if we define the upward direction as positive, the acceleration due to gravity will be negative. What follows are several example questions we could answer using this no-friction model for the motion of a falling object.

1. If I drop a ball from rest from a height $y=10\text{m}$, how long does it take to strike the ground at height 0m ?

$$s_f = s_i + v_i \Delta t + \frac{1}{2} a \Delta t^2$$

$$0 \text{ m} = 10 \text{ m} + \frac{1}{2} (-9.8 \text{ m/s}^2) \Delta t^2$$

$$\Delta t = \sqrt{\frac{-10 \text{ m} \times 2}{-9.8 \text{ m/s}^2}} = \sqrt{2.04 \text{ s}^2} = \pm 1.43 \text{ s}$$

It strikes the ground with speed:

$$v_f = v_i + a \Delta t$$

$$v_f = 0 + (-9.8 \text{ m/s}^2) 1.43 \text{ s} = -14 \text{ m/s}$$

2. If, instead, I throw the ball downward with an initial speed of 10m/s , how long does it take to reach the ground?

$$s_f = s_i + v_i \Delta t + \frac{1}{2} a \Delta t^2$$

$$0 \text{ m} = 10 \text{ m} + (-10 \text{ m/s}) \Delta t + \frac{1}{2} (-9.8 \text{ m/s}^2) \Delta t^2$$

$$(4.9 \text{ m/s}^2) \Delta t^2 + (10 \text{ m/s}) \Delta t + 10 \text{ m} = 0 \text{ m}$$

Notice that this is a quadratic equation with the familiar form:

$$ax^2 + bx + c = 0$$

which has solutions given by:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

which in this case are:

$$\Delta t = \frac{-(10 \text{ m/s}) \pm \sqrt{(10 \text{ m/s})^2 - 4(4.9 \text{ m/s}^2)(10 \text{ m})}}{2(4.9 \text{ m/s}^2)}$$

$$\Delta t = 0.73 \text{ s} \quad \text{or} \quad -2.8 \text{ s}$$

The solution we seek is the positive one, the one with $\Delta t = 0.73\text{s}$. At this moment the object has speed:

$$v_f = v_i + a \Delta t = (-10 \text{ m/s}) + (-9.8 \text{ m/s}^2)(0.73 \text{ s}) = -17.2 \text{ m/s}$$

Two points are important to make here. First, if you are uncomfortable with solving quadratic equations in this way, I again encourage you to review this math. You'll have to be able to do this over the coming months. Second, this equation, like every quadratic equation, has two different roots. In this case one is negative and one positive. The physical meaning of this is as follows:

- We asked the question: "If the motion of this object is governed only by the uniform acceleration of gravity, and if we know it passed a point $y = 10\text{m}$, traveling downward at 10 m/s at time $t = 0$, at what time did it pass the point $y = 0\text{m}$?"
- The answer we seek is the one which occurs *after* the ball is thrown

But it is perfectly possible for the object to have passed this point at an earlier time, while all the time executing motion described completely accurately by these equations. How does this happen?

These same conditions could be created by launching the ball from the ground ($y = 0\text{m}$) at a time equal to $t = -2.8\text{ s}$, with a speed equal to an opposite the speed we have calculated it will hit with. So if we launch a ball upward, with speed 17.2 m/s at time $t = -2.8\text{ s}$, it will pass the point $y = 10\text{m}$ at time $t = 0$, traveling down at -10 m/s , and strike the ground at time $t = 0.73\text{s}$, traveling down at -17.2m/s .

Notice that even in the first problem, in which I just dropped a ball from 10m up, there were really two solutions: we found that $\Delta t = \sqrt{2.0} = \pm 1.4$. We ignored the negative solution at that point, but it is mathematically allowed. In this case it just indicates that we could have thrown the ball up with a speed of 14 m/s at time $t = -1.4$. In either case we will have the motion from $t = 0$ to $t = 1.4$ be the same.

This emphasizes a basic symmetry in the motion of objects moving under constant acceleration. When I throw a ball up and down, there are two equal and opposite parts of the motion; a continual slowing as the ball rises, and a continual speeding up as it falls. So when I launch a ball in the air with speed $+10\text{ m/s}$, I know that it will return to the ground with speed -10 m/s . Once we have an understanding of this, we can use it to look at all problems involving constant acceleration.

3. If I launch a ball upward with speed 25 m/s , how long will it take it to come back down? What's the "easy" way to approach a problem like this? In the absence of air resistance, I know that it will reach the ground with a "final" speed of exactly -25 m/s , so I could use:

$$v_f = v_i + a\Delta t$$

$$(-25\text{ m/s}) = (25\text{ m/s}) + (-9.8\text{ m/s}^2)\Delta t$$

$$\Delta t = \frac{-50\text{ m/s}}{-9.8\text{ m/s}^2} = 5.1\text{ s}$$

How else could we figure this out?

- Use the fact that at the top of its flight $v=0$, and that this must be half-way through its entire flight
- Use the equation $\Delta s = v_i\Delta t + 1/2a\Delta t^2$, with $\Delta s=0$

Any of these methods would be correct. Often the key in physics is selecting the correct approach which admits easiest analysis of the problem. Very often the simplifying point involves taking advantage of some kind of symmetry in the problem.

Before we leave this topic, we should think carefully about what we've just done. We asked questions about the real motions of objects (balls thrown up and down) and analyzed what they would do if they moved without air friction. We did this because calculating the details of the motion is simple in this idealized model. But of course this model is not a perfect representation of reality. Much of the time it is not even a *reasonable* representation of reality. Any time we apply this approach to a case where the effects of air friction are not small, we know for sure that this model will give us answers which are very imprecise.

Imagine, for example, that the object we drop in case 1 above is a feather. Will this feather fall ten meters in 1.4 seconds? Will it strike the ground traveling at 14 m/s? Of course not! This is a fine example of a case where the 'no-friction' model is a terrible one; we should know we will go very far wrong if we use it. When will this no-friction model be reasonable? Any time the frictional forces which accompany the motion remain small in size compared to the weight of the object. Typically this will happen when an object is very massive (so that its weight is large) and moving relatively slowly (so that the frictional forces resisting its motion are small). Both things matter. For an object of low mass, like a feather, even a very slow motion can create friction which is large compared to its weight. For an object of large mass, a high enough velocity will always make the friction large compared to the weight. But so long as the object is reasonable massive and moving reasonably slowly, we can use the no-friction model and expect a reasonably accurate answer.

Another force calculation example

Let's put this no-friction falling model together with Newton's second law to create a quantitative model for a bouncing ball. Imagine that a 0.2 kg rubber ball is dropped from a 2 m height. When it reaches the ground, it reverses its direction, bouncing off the ground and traveling back up to about where it started. How could we estimate the force exerted on it by the floor?

First, what is the momentum change? When it reaches the floor it has been accelerated by gravity through 2m of distance. The final velocity it has can be estimated from one of our equations for object moving under constant acceleration:

$$v_f^2 - v_i^2 = 2a(s_f - s_i) \quad \text{or} \quad v_f^2 = 2(-9.8 \text{ m/s}^2)(0 \text{ m} - 2 \text{ m}) = 39.2 \text{ m}^2/\text{s}^2$$

$$v_f = -6.3 \text{ m/s}$$

After it hits the floor it will bounce back off with just about the same velocity it came in with. This is what enables the rubber ball to bounce back up to almost the same height it was released from. We know that the force exerted on the ball changed its momentum from downward and larger to upward and large. How did its momentum change during the collision with the ground?

$$\Delta \vec{p} = \vec{p}_f - \vec{p}_i$$

$$\Delta \vec{p} = (0.2 \text{ kg})(6.3 \text{ m/s})\hat{y} - (0.2 \text{ kg})(-6.3 \text{ m/s})\hat{y} = (2.6 \text{ kgm/s})\hat{y}$$

Notice that throughout we have to keep careful track of the fact that this is a vector!

OK, we have the momentum change. That's equal to the impulse, so:

$$\vec{F}\Delta t = (2.6 \text{ kgm/s})\hat{y}$$

What direction is the force in? It must be up, because that's the direction of the impulse. Is this surprising? Not really, because in order to make the ball change direction from moving down to moving up, the floor will have to push upward.

How large is the force? To estimate this we have to estimate the amount of time during which the force acts. Is it 1s? Is it 1/1000 of a second? It's probably somewhere in between, say 1/20 of a second. So, we would estimate that the force is about:

$$\vec{F}_{est} = \frac{\Delta\vec{p}}{\Delta t} \approx \frac{(2.6 \text{ kgm/s})\hat{y}}{(0.05 \text{ s})} = (52 \text{ kgm/s}^2)\hat{y}$$

So this force is about equal to the weight of 50 medium sized apples, a pretty sizeable impact.

What if we made the impact shorter, by making the ball harder for example? If we made the impact shorter, the force required to reverse the momentum of the ball would become correspondingly larger. What if we made it longer, by bouncing it off a trampoline for instance? In this case, the force required to reverse the motion of the ball could be much smaller. Such a small force can still achieve the same total change in momentum so long as it is able to act over a longer period of time.

This is how airbags, cushioned tennis shoes, and shoulder pads work. By allowing the momentum change to take place over a longer time, these devices enable smaller, less harmful forces to produce the same changes in momentum which would otherwise require quite large forces.

7.4 Not so free fall: small things falling slowly through the air

As we have seen, using the no-friction model for a falling object is likely to lead to errors. So let's try to improve this model, and analyze what would happen to an object dropped from rest which *was* subject to air friction. Recall that we have already introduced (in Chapter 5) models for the frictional force experienced by an object moving through a fluid; two different models in fact, one appropriate for small things moving slowly and the other for large things moving rapidly. We also already know what the end result of this motion will be; falling at a constant 'terminal' velocity. But what happens in between? How does the position and velocity change?

Let's first work out what the motion would be if the falling object were small, so that even when it began to fall it would never travel very fast. In this case, we might expect the small-slow friction to always apply. What does Newton's second law tell us about this motion?

$$\vec{F}_{total} = \vec{F}_{friction} + \vec{W} = m\vec{a}$$

In this case, all the motion is in the y direction, so we can drop the vector notation, and taking the upward direction to be positive, write:

$$-6\pi\eta r v_y - mg = ma_y$$

Notice that the weight is negative here because it acts downward. The frictional force has a minus sign in front of it because it always opposes the motion, it always acts opposite the direction of motion v . So if the object is moving upward, with v positive, the frictional force acts down. If the object is moving downward, with v negative, the frictional force acts upward. Recognizing that acceleration is the time rate of change of the velocity, this can be rewritten:

$$m \left(\frac{dv_y}{dt} \right) + 6\pi\eta r v_y + mg = 0$$

If we could find any function $v_y(t)$ which is a solution to this equation, it would be our prediction for the motion of this small, slow falling object. This equation is a first order linear differential equation. The solution to this equation would take the form:

$$v_y(t) = v_f (1 - e^{-\lambda t})$$

Does this work as a solution to this equation? To find out, we first calculate the derivative of this function $v(t)$, and insert both into the equation above and see what it tells us:

$$a_y(t) = \frac{dv_y(t)}{dt} = \lambda v_f e^{-\lambda t}$$

$$m \lambda v_f e^{-\lambda t} + 6\pi\eta r v_f (1 - e^{-\lambda t}) + mg = 0$$

Examining this equation, we see that it can be true if λ and v_f take on the values

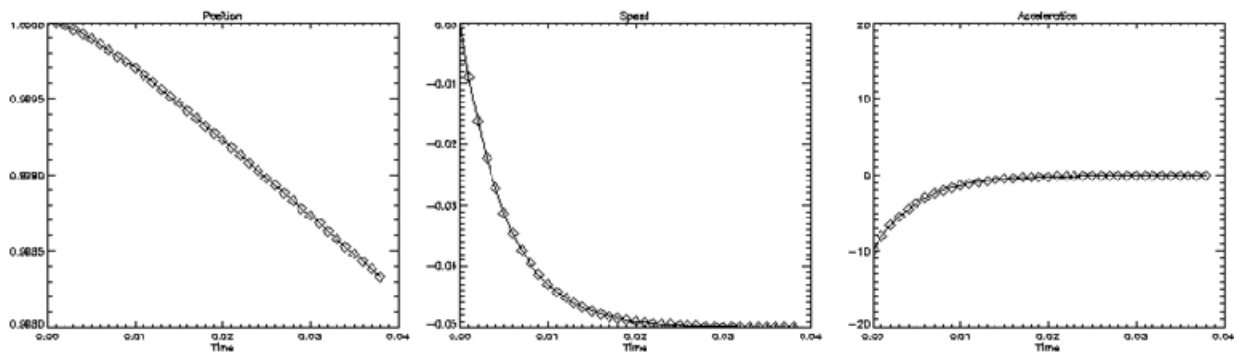
$$\lambda = \frac{6\pi\eta r}{m}$$

$$v_f = -\frac{mg}{6\pi\eta r}$$

Putting these into the equation makes the total prediction for the velocity of the object as a function of time:

$$v_y(t) = -\frac{mg}{6\pi\eta r} \left(1 - e^{-\frac{6\pi\eta r}{m} t} \right)$$

So a small-slow object will start out at rest with an initial acceleration $a_y(0) = g$. It will then gradually speed up, with an acceleration which falls toward zero and a velocity which increases in magnitude toward v_f . After a while, it will fall at a constant terminal velocity determined by the balance between the weight of the object and the frictional force which resists its motion through the fluid. This much we had already determined back in Chapter 5. The details of this kind of motion are shown in the position-time, speed-time, and acceleration-time plots below.



Let's recall what we've done. We specified a case where two forces act, the downward force of gravity and an upward force of small-slow fluid friction. We then wrote down Newton's second law for this case. At this point, the physics of the problem was done; this is all physics has to tell us about the problem.

Identifying a function $v(t)$ which would describe this motion is a mathematical problem. In this case we were able to identify an analytic form for the solution, something which is often not possible. As a last step, we inserted our proposed $v(t)$ into Newton's second law to determine the precise parameters in the general form we proposed.

In most realistic cases, finding an analytic solution for the equations of motion as expressed in Newton's second law is impossible. We will see an example in a moment. This does *not* mean that we cannot predict the motion of such an object. It just means that we can't write down the predicted motion in a simple equation, as an analytic function.

Remember, we are working here on small things which as a result fall slowly through the air. If the objects were large, they would have terminal velocities too large for this small-slow form of friction to be appropriate. How small is small here? For objects falling through the air, this analysis is most appropriate for things which are very small indeed: smaller than 1 millimeter; single celled ciliates, pollen grains, tiny water droplets which make up mist, dust grains, fungal spores, and viruses. For these tiny objects, terminal velocities are very small, much smaller than the typical flows of air we associate with breezes and temperature gradients in the air. As a result, such objects may never settle out of the air, but instead be carried aloft by large scale flows of air just as they try to fall out.

Not so free fall: large things falling rapidly through the air

With this detailed example under our belts, let's look at the other extreme, large objects which fall rapidly through the air. The basic analysis is the same, but because we're now looking at large-fast friction, the form of Newton's second law will be different.

$$\vec{F}_{total} = \vec{F}_{friction} + \vec{W} = m\vec{a}$$

Once again, the motion is all along the y axis, so we drop the vector notation and take the upward direction to be positive:

$$\frac{1}{2}C\rho Av^2 - mg = ma$$

$$m\left(\frac{dv}{dt}\right) - \frac{1}{2}C\rho Av^2 + mg = 0$$

There are a few differences between this and the equivalent equation for small-slow falling objects. First a detail; the sign in front of the frictional force here is positive, while we wrote it negative for the small-slow case. This is because small-slow friction is linearly dependent on velocity. When the velocity is negative, the frictional force is positive; hence the minus sign. With large-fast friction, the force is dependent on the square of the velocity. As a result, we cannot encode its direction within the equation, but must instead assign its sign according to what we know about the motion. Here the frictional force is positive, and we must account for that with a plus sign.

The more important difference is that this is no longer a first order linear differential equation. It is, instead, a non-linear differential equation; it contains the term v^2 . There is no analytic function $v(t)$ which is a solution to this equation. It is impossible to write down as a simple function describing the motion of such an object. What can we do in a case like this? Let's begin by looking back at what the physics tells us; what does Newton's second law say about this motion? Take the form we wrote for this above, and solve it for the time rate of change of the velocity:

$$\frac{dv}{dt} = \frac{C\rho A}{2m}v^2 - g$$

This equation should govern the motion of a large falling object, one which will end up with a large terminal velocity. It shows what we expect. At the first instant, when the object is released from rest, the downward acceleration is just that due to gravity. At some later time, the upward frictional force will balance the downward gravitational force, and the object will fall at a constant terminal velocity. How can we use this to predict the motion between those two limiting times?

Let's start at the beginning. When this large thing first starts to fall, its velocity is zero. So at this first moment, the acceleration is:

$$a(0) = \left.\frac{dv}{dt}\right|_{t=0} = -g$$

We know the initial velocity is zero, and the initial acceleration is $-g$. What will the motion be like a short time Δt later? Now the velocity will be

$$v(\Delta t) \approx -g\Delta t$$

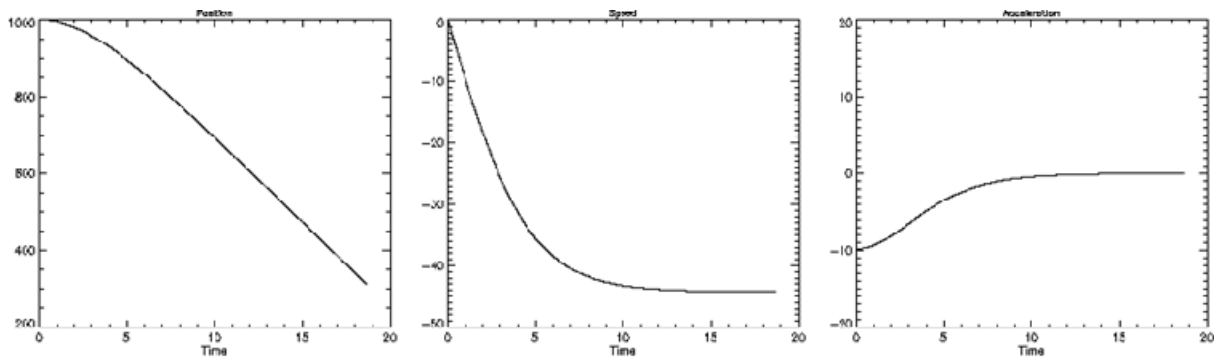
$$a(\Delta t) \approx \frac{C\rho A}{2m}[v(\Delta t)]^2 - g$$

After the next short time step these will be:

$$v(2\Delta t) = v(\Delta t) + a(\Delta t)\Delta t$$

$$a(2\Delta t) = \frac{C\rho A}{2m} [v(2\Delta t)]^2 - g$$

And so on. Given some initial conditions, we can predict the velocity and acceleration of this falling object at any point in the future. We can't write down the answer in a simple analytic form, but we can predict it with arbitrary accuracy. Figures showing the position-time, speed-time, and acceleration-time graphs for this motion are shown below. They look very like the ones developed above for small-slow falling objects, but they are different in significant ways. The way acceleration approaches zero is not exponential; it starts changing more slowly than it did in that case. Likewise, the approach of the velocity to its final terminal value is also not exponential. Neither can be described by a simple function, though they can be predicted in this numerical way.



Solving for the motion in this way is what we almost always have to do in physics. Very few real motions can be accurately expressed using analytic functions. But as long as we know what forces act on an object at each instant, we can *always* predict its motion accurately. This is the great triumph of Newtonian mechanics.

Falling (and rising!) through water

We've just examined, in some detail, the problem of falling through air. We looked at cases where both extremes of fluid friction applied: small objects falling slowly and large objects falling rapidly. In both cases, we considered only the downward force of gravity and an upward force due to fluid friction. There is another force which might act; the buoyant force. Ignoring this is reasonable when an object is in the air, because the density of solid or liquid objects is typically so much larger than that of the air, making the buoyant force negligible compared to the weight.

But when an object is submerged in water ignoring the buoyant force is rarely appropriate, and often leads to disastrously incorrect conclusions. To see this, let's consider an object released from rest half way between the bottom and the surface of a deep lake. We will take the case of a large object which will move rapidly through the water, allowing us to assume the large-fast friction. What are the forces which act on this object? There is an upward buoyant force, a downward weight, and a frictional force which resists the motion of the object.

$$\vec{F}_{\text{total}} = \vec{F}_{\text{buoyant}} + \vec{F}_{\text{friction}} + \vec{W} = m_{\text{object}} \vec{a}$$

$$\rho_{\text{water}} V_{\text{object}} \mathbf{g} + \frac{1}{2} C \rho_{\text{water}} A v_{\text{object}}^2 - \rho_{\text{object}} V_{\text{object}} \mathbf{g} = \rho_{\text{object}} V_{\text{object}} a_y$$

Here we have assumed that the object will sink downward, so that the resistive friction force will be upward. Once again, let's consider what happens just at the moment when the object is released (when its velocity is zero) and then again when the motion ceases to change (when it reaches terminal velocity).

At the first instant, when $v_{\text{object}} = 0$, we can rewrite the sum of forces as:

$$\rho_{\text{water}} V_{\text{object}} \mathbf{g} - \rho_{\text{object}} V_{\text{object}} \mathbf{g} = \rho_{\text{object}} V_{\text{object}} a_y$$

$$a_y = \left(\frac{\rho_{\text{water}}}{\rho_{\text{object}}} - 1 \right) \mathbf{g}$$

Now that the buoyant force can no longer be ignored, the initial acceleration is not just the downward acceleration due to gravity \mathbf{g} , it can take on any value, and be positive (if $\rho_{\text{object}} < \rho_{\text{water}}$) or negative (if $\rho_{\text{object}} > \rho_{\text{water}}$). Objects less dense than water will rise up through it when submerged. Objects more dense than water will sink down through it when submerged.

After this first moment, the object will begin to accelerate upward (or downward), and friction will begin to increase, lowering its acceleration. Eventually it will reach a terminal velocity, either rising or sinking, which we can find by setting the acceleration a_y equal to zero. Let's imagine first that it sinks, so that the frictional force acts upward.:

$$\rho_{\text{water}} V_{\text{object}} \mathbf{g} + \frac{1}{2} C \rho_{\text{water}} A v_{\text{terminal}}^2 - \rho_{\text{object}} V_{\text{object}} \mathbf{g} = 0$$

$$v_{\text{terminal}} = \sqrt{\frac{2}{CA} \left(\frac{\rho_{\text{object}}}{\rho_{\text{water}}} - 1 \right) V_{\text{object}} \mathbf{g}}$$

Since the density of most solid objects is not so different from that of water, these terminal velocities will typically be much smaller than the terminal velocities experienced by objects falling through the air. Not only is water a much more dense material to move through, but the buoyant force it provides can be quite substantial in comparison to the weight. At the end of this sinking, the object will come to rest on the bottom of the water, at which point it will be partially supported by a buoyant force, and partially supported by a normal force exerted by the bottom of the lake.

If an object is less dense than water, it will rise through it, reaching an upward terminal velocity which we can derive in the same way:

$$\rho_{\text{water}} V_{\text{object}} \mathbf{g} - \frac{1}{2} C \rho_{\text{water}} A v_{\text{terminal}}^2 - \rho_{\text{object}} V_{\text{object}} \mathbf{g} = 0$$

$$v_{\text{terminal}} = \sqrt{\frac{2}{CA} \left(1 - \frac{\rho_{\text{object}}}{\rho_{\text{water}}} \right) V_{\text{object}} \mathbf{g}}$$

When the object reaches the surface it will rise up partly out of the water, but just until the buoyant force supporting it is equal to its weight, and there it will float. How does this final state work out? Recalling that the buoyant force is always the weight of the fluid displaced, we can write:

$$W_{\text{displaced}} - W_{\text{object}} = 0$$

$$\rho_{\text{water}} V_{\text{object}}^{\text{in water}} g - \rho_{\text{object}} V_{\text{object}}^{\text{total}} g = 0$$

$$\frac{V_{\text{object}}^{\text{in water}}}{V_{\text{object}}^{\text{total}}} = \frac{\rho_{\text{object}}}{\rho_{\text{water}}}$$

Such floating objects rest there, partly submerged, so that the fraction of their volume which is submerged is just determined by what fraction their density is of the density of water. People are almost as dense as water, especially when they breathe out, and hence float almost entirely in the water. Beach balls, by contrast, are much less dense than water, and hence float almost entirely out of the water. Once you know this, a quick glance should allow you to estimate the density of a floating object relative to water.



7.5 Summing up motion in one dimension: speeding up and slowing down

We have been considering cases in which unbalanced forces act only along or opposite the direction of motion. This includes all kinds of cases of speeding up and slowing down while traveling straight. As always, predicting what will happen in this kind of motion requires knowing what the forces are. If you know the net force on an object, you can use Newton's second law to predict its acceleration. Once you know the acceleration, you can use the kinematic relations between acceleration, velocity, and position to predict exactly how it will move. So, if you know the forces, you know what will happen.

We discussed two special cases as well. One was motion under the influence of constant acceleration. The position, velocity, and acceleration as a function of time have very simple analytic forms for this situation. This makes it a nice model for those parts of a motion in which the total force acting on an object is approximately constant. As an example of this we looked at the motion of a ball dropped or tossed upward in the air. While this constant acceleration model is usually not precise (because net forces are rarely exactly constant) it is a useful first approximation in many cases. This sort of 'start with a basic model' should bring to mind the spherical cow we used to explore scaling relations at the start of the class.

As an example of something more realistic, we looked at the case of falling through the air when we acknowledge that friction is present. In this case, an object released from rest starts out feeling a

downward force equal to its weight. This causes it to accelerate downward. As its speed increases, the frictional force resisting its downward motion gradually increases. As a result, the *net* downward force is gradually reduced, and the downward acceleration decreases with it. This process continues until finally the upward frictional force just balances the downward weight. At this point, the net force is zero, and the object continues to fall at a constant terminal velocity.

Fortunately, you rarely experience this precise kind of terminal velocity motion. It usually ends badly...But most of the motions you do experience are very like this. You start from rest, and a force applied to you creates acceleration. This acceleration continues for a while, gradually becoming smaller as the forces which resist your motion build up, until at last the forces pushing you forward are just balanced by those pushing you backwards. From then on your velocity remains constant. What are some examples of this?

- Walking: you push backward on the floor with your foot, it pushes forward on you, and you start to accelerate. That forward force is pretty quickly balanced by backward forces which come (mostly) from the landing of the foot you put forward on the ground. The average force acting on you becomes zero, and you walk along at constant velocity.
- Running: this is similar, except that you reach high enough speeds for friction with the air to make a significant contribution to the force resisting your forward motion.
- Bicycling: By riding a bicycle, you reduce the ground friction resisting your forward motion quite substantially. As a result, you can travel much faster, and your final speed is limited by air friction more than friction with the ground. Think about what you do when you cycle on level ground. You pedal just hard enough to keep going at constant speed. To go faster, you adjust how hard you pedal to match the resistive force at that particular speed. Going up or down a hill you have an additional force from gravity either resisting or aiding your motion.
- Driving a car: Again, in this case you can adjust the force propelling the car forward. You can make it overcome the resisting force and increase the speed, balance the resisting force and travel at constant speed, or be smaller than the resisting force and slow down. Here the changes are really obvious, as they are controlled by just pushing harder or letting up on the gas pedal. You can't travel along at 70 mph without keeping your foot on the gas. You must be pushed forward with a force equal to the friction which is resisting your motion.

Notice that in all these cases the steady final velocity you achieve involves feedback. In cycling, you pedal just hard enough to counteract the resistive force. If you want to travel faster, you pedal harder and speed up until the resistive force matches this new, stronger, forward force. If you're going faster than you want to, you ease up on pedaling and let the frictional force (now larger than the forward force) slow the bike down.

In starting normal forward motion your feet (on the bike) or the motor (in the car) makes the wheels push backward against the ground, trying to slip over it. Static friction between the wheel and the ground then pushes back, driving the wheel (with the bicycle and rider) forward. When you want to stop, you do the opposite, using your brakes to slow the wheel's rotation. This would tend to make the wheels skid over the ground. So the wheel pushes forward on the ground, and the frictional force with the ground pushes backward on the wheel, slowing the wheel (and the bicycle and rider).

In all these cases the ability to *actively control* the forward and backward force which acts on you is what allows you to control your motion, speeding up and slowing down at will. The source of this active control is friction, and in particular the passive, static friction form of friction. Because it is static, it has an adjustable value, and you can control it.

This is in contrast to what happens when this interaction between your wheels and the ground is no longer static friction, but kinetic friction. Since kinetic friction is an active, rather than a passive force, its magnitude is fixed; you lose the ability to control the forces on you. If you start to slip on ice, then the frictional force between you and the ice becomes a fixed kinetic friction value. Nothing you can do changes this, so you lose control of your motion and can only continue to skid until you come to a stop.

It is interesting to investigate how long it takes an organism to reach full speed. This will be dependent on the maximum acceleration each creature might achieve, and it is worth considering how acceleration might scale with size. We know that acceleration is related to force and mass through:

$$a = \frac{F}{m}$$

We have seen before that available muscle force typically scales with size^2 , while mass scales with size^3 , so we might expect the accelerations organisms can achieve to be larger for smaller creatures, and smaller for large ones.

Indeed the very largest accelerations observed in life are for tiny things, like the spores of fungi, which may be launched with accelerations of around 10^6 m/s^2 ; one hundred thousand times the acceleration due to gravity. Some insects take off with accelerations of 10^4 m/s^2 , which you probably know if you have tried to catch a grasshopper: once she starts moving, she's gone before you know it. Mammals and most creatures on our scales are limited to accelerations less than about ten times the acceleration due to gravity, though it is not clear whether this is primarily limited by the strength of our muscles or our inability to survive much larger accelerations. In any case, though small things may not travel fast, they get up to speed very much more quickly than we do.

A Quick Summary of Some Important Relations

Momentum and its changes:

Momentum measures the vector quantity of motion for an object. Changes in it are found from the vector difference between final and initial momentum.

$$\vec{p} = m\vec{v} \qquad \Delta\vec{p} = \vec{p}_{\text{final}} - \vec{p}_{\text{initial}}$$

Force and the rate of change of momentum:

Net force determines the time rate of change of the momentum.

$$\vec{F}_{\text{net}} = \frac{d\vec{p}}{dt} \qquad \vec{F}_{\text{net}}^{\text{average}} = \frac{\Delta\vec{p}}{\Delta t}$$

Impulse and change in momentum:

Impulse is the sum of the force at each instant multiplied by the time over which it acts. This quantity determines the total change in momentum.

$$\text{Impulse} = \int_{t_{\text{initial}}}^{t_{\text{final}}} \vec{F}(t) dt = \Delta\vec{p} \qquad \vec{F}_{\text{net}}^{\text{average}} \Delta t = \Delta\vec{p}$$

Force and acceleration:

For most purposes, the momentum of an object changes only by changing its velocity. In these cases we can rewrite Newton's second law as:

$$\vec{F}_{\text{net}} = m \frac{d\vec{v}}{dt} = m\vec{a}$$

8. Turning the corner: dynamics and motion in 2 and 3 dimensions

- 1) Moving through the real world; motion in two and three dimensions
 - i. Basic observations from kinematics
 - ii. Position vectors and their change in time
 - iii. Extracting velocity and acceleration from these
 - iv. When you see the accelerations, you know the total forces
- 2) A first useful model: circular motion at a constant speed
 - i. The details: position velocity and acceleration
 - ii. Two important instances: circular orbits and the centrifuge
 - iii. Using this as an approximation for more complex curved paths
- 3) Starting and stopping rotation
 - i. Torque and angular momentum
 - ii. Rotational inertia of 'rigid bodies'
 - iii. The parallel between linear and rotational motion
- 4) A second useful model: motion with a single, constant force
 - i. The ideal projectile and superposition in physics
 - ii. Details of ideal projectile motion
 - iii. Newtonian dynamics and boundary value problems
- 5) More complex cases: projectiles with friction and beyond
 - i. When forces are known, Newton's laws still do it all
 - ii. Numerical calculation of paths: projectiles with friction
 - iii. Numerical calculation of paths: orbits of planets
 - iv. Other examples of numerical calculations: fluid dynamics and the growth of cosmic structure
- 6) The legacy of Newtonian dynamics
 - i. Mathematical models built on simple principles
 - ii. Reductionism in modern science

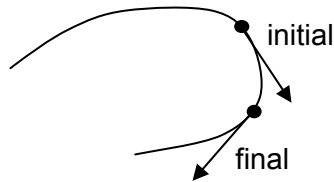
Physics for the Life Sciences: Chapter 8

8.1 Moving through the real world; motion in two and three dimensions

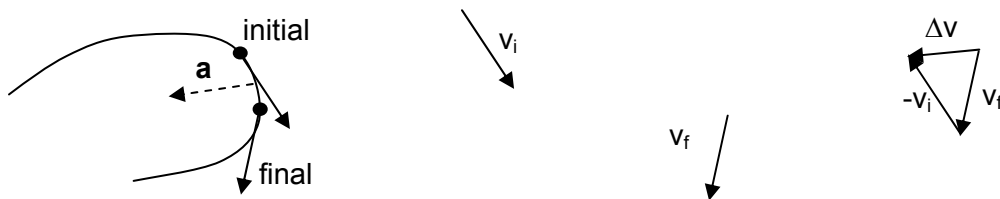
While motion in one dimension captures many interesting cases, it leaves out the rather important fact that we live in a three-dimensional space. We need to understand not only how an organism might speed up and slow down, but also how it might change direction. Indeed the most general case involves a combination of speeding up or slowing down (changing the magnitude of the velocity) and turning (changing the direction of the velocity). Just because it's simpler to draw, we're going to talk mostly about two-dimensional motion, rather than the full 3D. The extension to motion in three dimensions instead of two is relatively easy to see.

Basic observations from kinematics

There are some very powerful things purely descriptive kinematics can tell us about absolutely any 2D motion. Consider motion along a simple curve at constant speed (constant magnitude of the velocity):



The velocity is changing in this motion. Its magnitude stays the same, but the direction is changing all the time. The velocity, as we have several times stressed, is always along this path of motion. In what direction is the acceleration? To find the acceleration we can draw the velocity vector at two successive moments, and find the vector change in velocity $\Delta\vec{v} = \vec{v}_f - \vec{v}_i$



The acceleration vector is

$$\vec{a} = \frac{d\vec{v}}{dt} = \lim_{\Delta t \rightarrow 0} \left(\frac{\Delta\vec{v}}{\Delta t} \right)$$

Note that acceleration is a vector in the direction of $\Delta\vec{v}$. So for the curved motion shown here, the acceleration is directed into the curve. In fact it has to be directed into the curve. If you think about it, the velocity is always turning into the curve, so that's always the direction of its change.

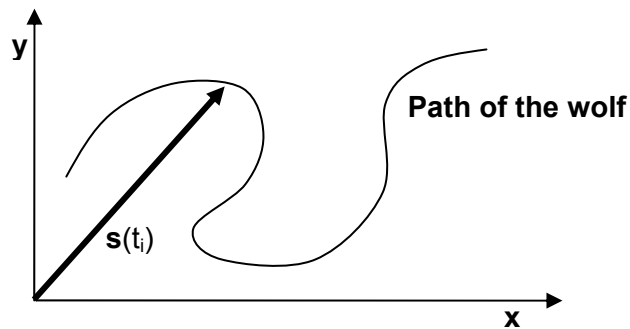
How large is the acceleration? For motion at constant speed along such a path, the size of the acceleration depends on how suddenly the path changes direction. This can be expressed in terms of the "radius of curvature"; the sharper the curve, the smaller this radius, and the larger the associated acceleration.

What if there is acceleration along the curve (speeding up or slowing down) as well as perpendicular to it (changing the direction)? The total acceleration will then have a component along the curve as well as a component into the curve (like the one shown above), so that the net acceleration is the vector sum of these two components. But no matter what, if there is curvature in the path of a moving object, there is some non-zero component of the acceleration toward the center of the curve.

Position vectors and their change in time

How do we find the acceleration associated with a particular path? Let's imagine you know (from a video for example) exactly where a wolf is at every moment while it chases a deer. This means you know its

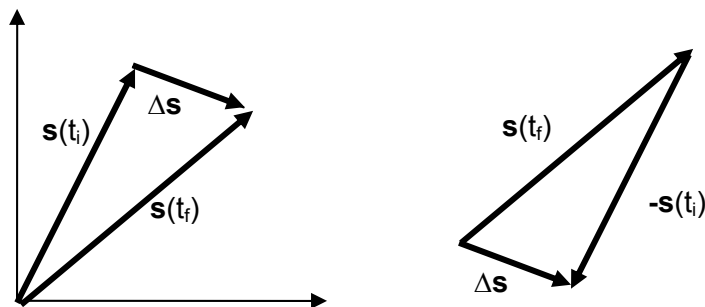
entire path $\vec{s}(t)$. Notice that we wrote the wolf's position at each instant as a vector \vec{s} . At each moment the animal is at a particular point in three dimensions (x , y , and z for example). The "position vector" $\vec{s}(t)$ is a vector which goes from the origin of the coordinate system to the location of the wolf at a particular moment t .



As the wolf moves from one position to another, the change in position is a displacement vector given by the equation:

$$\Delta\vec{s} = \vec{s}(t_f) - \vec{s}(t_i) \quad \text{or} \quad \vec{s}(t_i) + \Delta\vec{s} = \vec{s}(t_f)$$

And can be shown in a figure like this:



Not surprisingly, the change in position is just a vector which goes from the initial position to the final position.

Extracting velocity and acceleration vectors from position as a function of time

Knowing the path of an object, its position as a function of time, it is always possible to find the displacement vectors corresponding to any time interval Δt . Given the displacement $\Delta \vec{s}$ for this little time interval, you can estimate the velocity during this period:

$$\vec{v}_{est} = \frac{\Delta \vec{s}}{\Delta t}$$

This equation should remind you that the direction of the velocity is always the same as the direction of the displacement $\Delta \vec{s}$: the velocity is always along the motion.

Tracking the changes in velocity from moment to moment allows you, in a similar way, to find the accelerations:

$$\vec{a}_{est} = \frac{\Delta \vec{v}}{\Delta t}$$

Since the *change in velocity* is not always in the direction of the velocity, the acceleration is *not* always in the direction of motion. This is true for the curved motion described above. The velocity is always along the direction of motion, but the acceleration may be partly or even completely perpendicular to this.

When you know accelerations, you know the total force which acts

Any time you see the whole motion of an object you can write down its path $\vec{s}(t)$. From this you can work out both the velocity $\vec{v}(t)$ and acceleration $\vec{a}(t)$ as a function of time. Once you know the acceleration at every moment, you can deduce exactly the size of the total force acting on it at each instant. This is a very important point; when you observe the motion of an object, you know what total force acts on it.

This is not an entirely new idea, in fact we have been using it extensively. In earlier chapters, we have repeatedly used the notion that when an objects motion does not change, when its acceleration $\vec{a}(t)$ is zero, the net force on it must be zero. Now we are simply applying the same notion to cases where the acceleration is not zero. In these cases, we know that:

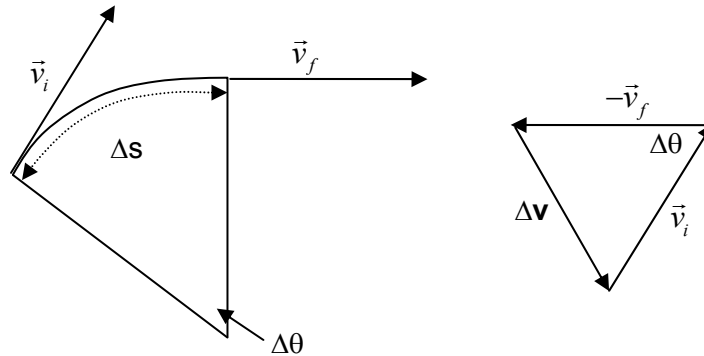
$$\vec{F}_{\text{Total}}(t) = m\vec{a}(t)$$

Most of the motions we see in the world, like a wolf chasing a deer or a chickadee flitting from tree to tree, are complex, with accelerations that change in complex ways. When a motion is complicated like this, the forces which create it must be complicated, changing in complex ways as the object moves.

There are, however, a couple of cases where curved 2D motion remains relatively simple: circular motion at constant speed, and motion under the influence of a constant force. As usual, we'll use these simple cases as starting models for describing and understanding more complex motions.

8.2 A first useful model: circular motion at a constant speed

What is the acceleration of an object traveling in a circle at a constant speed? Let's first work this out using a simple geometric approach. Consider the following diagram:



In the diagram on the right we have used the fact that if the angle $\Delta\theta$ is the one shown at the center of the circle, it is also the angle between the two vectors \vec{v}_i and \vec{v}_f . This is so because the radius of the circle is always perpendicular to the velocity vector \vec{v} . The acceleration $\vec{a} = d\vec{v}/dt$, and to determine it we need to be able to express the change in velocity $\Delta\mathbf{v}$ in a simple way. The angle $\Delta\theta$, measured in radians, is defined as:

$$\Delta\theta = \frac{\Delta s}{r}$$

Now if the angle $\Delta\theta$ is small, as it would be if we were considering the motion through a short period of time Δt , we can solve for the angle $\Delta\theta$ by examining the little triangle on the right. In this triangle, the side opposite $\Delta\theta$ has length Δv . This takes the role of the arc length Δs in the definition of the angle given above. The magnitude of the velocity vector $|\vec{v}|$ takes on the role of the radius, and we can write $\Delta\theta$ as:

$$\Delta\theta = \frac{\Delta s}{r} \approx \frac{\Delta v}{v}$$

Where the symbol v refers to the magnitude of the velocity, which isn't changing. We can use this to find the desired magnitude of the change in velocity:

$$\Delta v = \frac{v\Delta s}{r}$$

but

$$a_{est} = \frac{\Delta v}{\Delta t} = \frac{v\Delta s}{r\Delta t} = \frac{v^2}{r}$$

In this equation, we have used the fact that $v = \Delta s / \Delta t$.

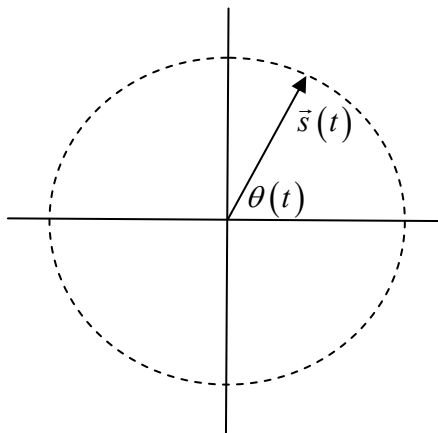
What direction is this acceleration in? As we noted before, in any curved path the acceleration will be into the curve. That's what's needed to bend the path of the object. In this case there is no acceleration along the motion (it doesn't speed up or slow down along the circle), so the acceleration must be perpendicular to the motion and directed toward the center of the circle. What does this mean?

Any time you see an object moving in a circle at a constant speed, you know its acceleration must have a size given by v^2/r , and it must be directed towards the center of the circle. This acceleration is called 'centripetal' or center-seeking acceleration. Whenever an object accelerates, you know that the net force on it must be non-zero. As a result, any object traveling in a circle at a constant rate must be experiencing a net force pointing toward the center of the circle with a magnitude of:

$$F_{\text{centripetal}} = ma_{\text{centripetal}} = \frac{mv^2}{r}$$

A more direct derivation

Let's use the ideas of the previous section more directly to examine the motion of an object traveling in a circle at a constant speed. We begin by describing the motion with a position vector $\vec{s}(t)$ pointing from the center of the circle to the location of the object at time t . How does this vector change with time? The length of this position vector never changes, it is always just the radius of the circle r . The position at any moment has x and y components which can be written in terms of the fixed radius r and the changing angle $\theta(t)$.



We specified an object traveling at constant speed; this implies that the angle increases at a constant rate, so that we can write:

$$\begin{aligned}\theta(t) &= \omega t \\ \vec{s}(t) &= r \cos(\theta(t))\hat{x} + r \sin(\theta(t))\hat{y} \\ \vec{s}(t) &= r \cos(\omega t)\hat{x} + r \sin(\omega t)\hat{y}\end{aligned}$$

Given this position as a function of time, we can find the velocity as a function of time in the usual way:

$$\vec{v}(t) = \frac{d\vec{s}(t)}{dt} = -r\omega \sin(\omega t)\hat{x} + r\omega \cos(\omega t)\hat{y}$$

When we started, we said this should be motion at constant speed, so that the magnitude of this velocity should be constant. Is this true, and what is this constant speed?

$$\begin{aligned}|\vec{v}(t)| &= \sqrt{v_x^2 + v_y^2} = \sqrt{r^2\omega^2(\sin^2(\omega t) + \cos^2(\omega t))} \\ |\vec{v}(t)| &= r\omega\end{aligned}$$

So yes, the speed (the magnitude of the velocity) is indeed constant. Now that we know the velocity, we can likewise find the acceleration:

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt} = -r\omega^2 \cos(\omega t)\hat{x} - r\omega^2 \sin(\omega t)\hat{y}$$

Checking the magnitude of the acceleration in this same way, we find that it is also constant in time:

$$\begin{aligned}|\vec{a}(t)| &= \sqrt{a_x^2 + a_y^2} = \sqrt{r^2\omega^4(\cos^2(\omega t) + \sin^2(\omega t))} \\ |\vec{a}(t)| &= r\omega^2\end{aligned}$$

If we like, we can rewrite this result in terms of the magnitude of the velocity, so that:

$$\begin{aligned}|\vec{a}(t)| &= r\omega^2 = \frac{|\vec{v}(t)|^2}{r} \\ a_{\text{circular motion}} &= \frac{v^2}{r}\end{aligned}$$

What can we say about the direction of the acceleration derived from this derivation? It is helpful to note that

$$\vec{a}(t) = -r\omega^2\vec{s}(t)$$

The acceleration vector is always opposite the direction of the position vector. Since the position vector always points from the center of the circle to a point on this circle, the acceleration always points toward the center of the circle, just as we found from our more geometric derivation above.

Implications of circular motion; turning corners

Let's consider some cases in which objects travel in circles and consider the origin and size of the centripetal forces needed to make them happen. First, a familiar case for the modern University student: rounding the corner of a highway entrance ramp while traveling at 30 miles per hour, or 13.4 m/s. A typical onramp has a radius of about 75 m, so this motion requires a centripetal acceleration of about:

$$a_{\text{centripetal}} = \frac{v^2}{r} = \frac{(13.4 \text{ m/s})^2}{75 \text{ m}} = 2.4 \text{ m/s}^2$$

To accelerate in this way, the car rounding the corner, and each object traveling with it, must experience a centripetal force large enough to make this happen. What is the origin of the force on the car, or on you as a passenger, or on the fuzzy dice hanging from your rear view mirror?



The force on the car comes from the interaction between its tires and the ground. When you turn your steering wheel to the right, the tires turn that way, this makes them push against the ground with a frictional force to the left and forward. The ground pushes back with a force to the right and backward. That force, transmitted through the frame of the car, provides the centripetal force which pushes the car around the corner.

When you sit in your car traveling straight forward at a constant speed no force is required; your motion is not changing. But when the car rounds the corner, your motion must change if you are to remain in your seat. To keep you moving with the car, a centripetal force large enough to create the required centripetal acceleration must be present. For you, sitting in your seat, this force may come from friction with your seat, the restraining forces applied by your seat belt, and sometimes from the doors of the car.

The fuzzy dice hanging from the rear view mirror must also receive a centripetal force if they are to remain with the car. Unlike the car or the passengers, the dice are pulled toward the center by the same string that supports them from the mirror. When this string pulls the dice to the side, they hang at an angle, something which can only happen when they apply both horizontal and vertical components of force to the dice. If the car is going to round the corner, if you are to stay in it, and if the dice are to come along, all must receive the appropriate centripetal force. If a large enough force is not available, it will be impossible to follow this path. Let's consider what this means for the car.

The force between the car's tires and the road is a static frictional force. The tires, when working as they should, do not slip over the road; their surface is instead laid down on the road as the tire arrives at a point, then lifted up again as the wheel moves off. Static friction has a fixed maximum value given by the relation:

$$F_s^{\max} = \mu_s F_{\text{normal}}$$

This static frictional force supplies the centripetal force needed to turn the car around the corner. Since the frictional force has a maximum value, there is a maximum centripetal force it can provide. Balancing the force available with the force required yields a constraint on how fast the driver can go around a turn with a given radius:

$$F_s^{\max} \geq F_{\text{centripetal}} \quad \text{or} \quad \mu_s F_{\text{normal}} \geq \frac{mv^2}{r}$$

$$\mu_s mg \geq \frac{mv^2}{r} \quad \text{or} \quad \mu_s g \geq \frac{v^2}{r}$$

This constraint tells us the maximum safe velocity for driving around a curve with a radius r will be given by $v_{\max} = \sqrt{\mu_s gr}$. Tires typically have a large coefficient of static friction on dry pavement, something like 0.7 for a typical tire with treads. For our example on ramp with a radius of 75 m, this implies a maximum exit speed of $v_{\max} = 23 \text{ m/s} = 51 \text{ mph}$.

What would happen if you tried to round this corner traveling faster than this? Static friction would no longer be able to provide enough centripetal force. Two things happen then. First, without enough centripetal force, the car cannot continue to drive on the circle. There is still some force turning it, just not enough to turn it sharply enough to stay on the circle. So it will still turn, but will exit the nice circular ramp. Second, since static friction is overcome, the tires will start to slide. This is especially bad, as at this point the friction becomes kinetic friction, an active force which exerts a basically uncontrollable constant force on the car. Now the driver cannot as easily adjust the forces which are moving the car, making it difficult to regain control.

There are, of course, other complications. The coefficients of friction are quite dependent on the details of the tires and road surfaces, which are especially sensitive to being wet or icy. Really wet roads can reduce the coefficient of static friction to 0.1 or less, and icy roads are worse still. This reduction implies a similar reduction in the maximum safe speed.

Road designers are aware of these problems, and often design highway ramps with a tilted road surface. This allows part, or in the right conditions all, of the centripetal force to come from the normal force between the tires and the ground. This force is always present, capable of being very large indeed, and independent of the conditions of the road or tires.

Implications of circular motion: planetary orbits and the universal law of gravity

Newton wanted to understand the motions of the planets. He knew several things about their orbits, mostly from the observations of earlier scientists like Tycho Brahe and Johannes Kepler. First of all, planets travel in very nearly circular orbits around the Sun, at very nearly constant speeds. Being the inventor of mechanics, he knew this required a force pulling them toward the center of the circle with a magnitude:

$$F_{\text{toward Sun}} = \frac{m_{\text{planet}} v_{\text{planet}}^2}{r_{\text{planet orbit}}}$$

Kepler had also discovered a close tight connection between the periods Γ and radii of planetary orbits. In particular, he found that

$$\frac{\Gamma_{\text{planet}}^2}{r_{\text{planet orbit}}^2} = \text{constant}$$

Since these orbital periods are related to the velocity and radii of the orbits, we can rewrite this as:

$$\Gamma_{\text{planet}} = \frac{2\pi r_{\text{orbit planet}}}{v_{\text{planet}}}$$

$$\frac{\Gamma_{\text{planet}}^2}{r_{\text{orbit planet}}^3} = \frac{(2\pi r_{\text{orbit planet}})^2}{(v_{\text{planet}})^2 r_{\text{orbit planet}}^3} = \frac{(2\pi)^2}{v_{\text{planet}}^2 r_{\text{orbit planet}}} = \text{constant}$$

$$v_{\text{planet}}^2 r_{\text{orbit planet}} = \text{constant}$$

Taking this result, based essentially on Kepler's observations of how planets really move, and combining it with his own understanding that circular motion requires a very particular centripetal force, Newton was able to write:

$$F_{\text{toward Sun}} = \frac{m_{\text{planet}} v_{\text{planet}}^2}{r_{\text{planet orbit}}} = \frac{m_{\text{planet}} v_{\text{planet}}^2 r_{\text{planet orbit}}}{r_{\text{planet orbit}}^2} = \frac{m_{\text{planet}} \text{constant}}{r_{\text{planet orbit}}^2}$$

In short, he was able to show that the force of gravity which holds planets in their orbits around the sun must obey a force law of the form:

$$F_{\text{toward Sun}} \propto \frac{m_{\text{planet}}}{r_{\text{planet orbit}}^2}$$

Extension of this idea to both the orbit of the moon around the Earth and the attraction of the Earth for objects falling toward it led Newton to the final form of his universal law of gravity, which expresses the gravitational interaction between any two objects with masses m_1 and m_2 as:

$$F_{\text{gravitational}} = \frac{Gm_1m_2}{r^2}$$

Implications of circular motion: the centrifuge

We have seen that an object traveling in a circle must experience a centripetal force to do so. This fact is regularly put to use in a device called a centrifuge, used to sort materials in solution according to their density. In a typical centrifuge a set of small vials are attached to a circular ring which then spins at a high rate. To turn so rapidly in a small circle a very large centripetal force must be applied to each tube. In fact, everything which rotates, including the contents of the tube, must also experience a large centripetal force to continue traveling in a circle.

Imagine a particle with density slightly larger than the fluid around it suspended in one of these tubes. When the centrifuge spins a larger centripetal force is required to keep the particle traveling in a circle than would be required for the fluid which might replace it. This situation is exactly parallel to the situation with gravity and the buoyant force. In that case, a force is required to hold up both the fluid and the object against the pull of gravity. If the object is more dense, it will sink through the fluid. If the fluid is more dense, it will flow below the object. In the centrifuge a force is required to keep both the fluid and the object traveling in a circle. If the object is more dense, it will ‘sink’ through the fluid toward the outside of the circle. If the fluid is more dense, the object will ‘float’ up toward the center of the circle.

Imagine a centrifuge which spins vials so that the fluid rotates around a center at a distance $r_{\text{centrifuge}}$ and rotates all the way around N times per second. The velocity of the sample will then be:

$$v_{\text{sample}} = 2\pi rN$$

and the centripetal acceleration of the sample would be:

$$a_{\text{centripetal}} = \frac{v^2}{r} = \frac{(2\pi rN)^2}{r} = r(2\pi N)^2$$

For typical laboratory centrifuges, the radius might be 20 cm and the rate of rotation 50 revolutions per second. A system like this will require a centripetal acceleration of 20,000 m/s^2 . This is about 2000 times the acceleration due to gravity. Such a system will allow slightly denser materials to sink through those less dense much faster than they would simply under the influence of gravity. In this sense, the centrifuge

is very like a system in which we create increased gravity, which then acts as ordinary gravity would to separate these materials in suspension.

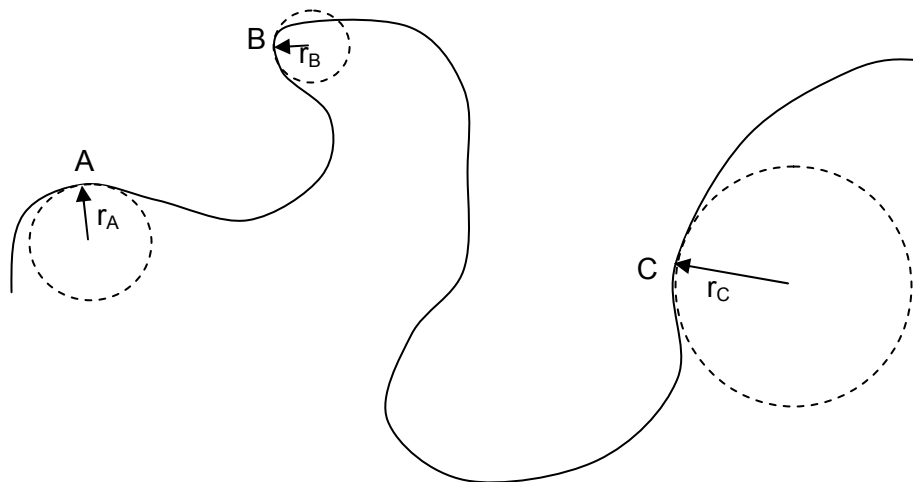
This ability to emulate the effects of large accelerations is used in pilot training as well. For this purpose room-sized centrifuges swing a person around in a circle to create accelerations ten and twenty times larger than ordinary gravitational acceleration. This is done because pilots and astronauts may experience such large accelerations while flying through sharp turns, and experiencing them first in a safe, controllable environment has proven an important safety step.

Using uniform circular motion as a model for more complex motions

There are two important comments to make about this. First, you might have guessed that the acceleration would be towards the center of the circle. If it was not *always* perpendicular to the motion, the object would either speed up or slow down along its direction of motion. Second, you can use this to approximate the acceleration perpendicular to the motion for any object traveling in a curve, especially when it's traveling at approximately constant speed. Just estimate the "effective radius of curvature" r_{eff} of the curve at that point and calculate:

$$a_{\perp} = \frac{v^2}{r_{eff}}$$

Note though, this won't tell you about acceleration *along* the curve. How does this work? Consider the picture drawn below:



Imagine this is the path of a horse which gallops along a winding road at constant speed. At each of the three points (A, B, and C) shown, the direction of the horse is changing, so there must be an acceleration. How large is each? Since the speed *along* the path is always the same, we'll just call that v . At point A, the horse is turning with a "radius of curvature" r_A , so the acceleration it experiences is approximately v^2/r_A . You can similarly estimate the acceleration at B and C.

Notice that when the radius of curvature is small, the horse is turning quickly. This quick turn requires a larger acceleration. You can see this from the equation because when it turns suddenly (as at B), the radius of curvature is small, and the resulting acceleration must be large. To have this larger acceleration, there must also be a larger force.

This ought to be a familiar fact. When you want to turn a corner sharply, you need a lot of force. You can do it if you wear good, sticky sneakers and push off really hard against the floor. But you can't do it if large frictional force you need is not available, perhaps because you're wearing socks on a hardwood floor, or because you're are trying to turn your car on an icy road. When a large force is not available, you can't turn a corner very sharply.

8.3 Rotating objects and rotational dynamics

Uniform circular motion might seem a relatively rare occurrence, relevant for planetary motion perhaps, and a useful approximation for parts of the complex paths objects might take. There is one additional way it appears, and this is perhaps more common; the rotation of more or less solid objects. When an object rotates around its center, each part of it travels in a circle. To do this, each part must experience a non-zero net force, acting toward the center of the circle. We know from our earlier work exactly how large this force has to be. Let's see how this works in a simple case.

Perhaps you have encountered a playground merry-go-round at some point in your childhood. This is a big disk with handles which spins around its center. Children with strong stomachs can get on and ride in circles. Several forces act on each child, including their weight, balanced by a normal force provided by the disk, and an unbalanced centripetal force provided by the handles, which pull each child inward toward the center of the disk. If this centripetal force were not there (if, for example, a child let go) they would no longer continue to travel in a circle but would instead continue moving in a straight line, flying off in a path tangent to the disk's edge, at least until they struck the ground and friction brought them to a halt.



Let's work out how large a centripetal force is required to keep a child traveling in a circle with the disk. Since the whole disk spins together, we can most easily describe its motion with an angular velocity that measures how large an angle the disk rotates through in each second. If the disk has a period of rotation Γ , we would say the angular velocity is:

$$\omega = \frac{2\pi \text{ radians}}{\text{Period of rotation}} = \frac{2\pi \text{ radians}}{\Gamma \text{ second}}$$

This symbol, the Greek letter ω , is usually used to represent an angular velocity. What are the units of angular velocity? Remember that radians are unitless; they are defined as a ratio of arclength to radius (length over length). So the units of angular velocity are really inverse seconds, or Hertz. But to keep track of what we're talking about, we will try to record angular velocities as having units of radians per second.

To find the centripetal force required to keep the child traveling in a circle we need to relate the angular velocity of the disk to the linear velocity of the child. Let's imagine the child rides a distance r_{child} from the center of the disk. Each time the disk spins, the child travels once around its full circumference, so we can write:

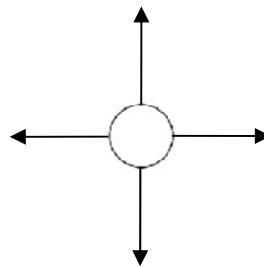
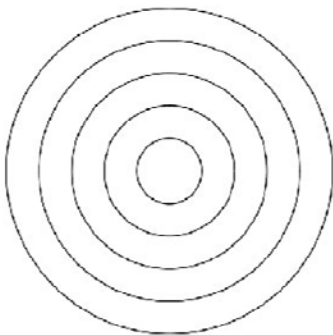
$$v_{\text{child}} = \frac{\text{Circumference}}{\text{Period of rotation}} = \frac{2\pi r_{\text{child}}}{\Gamma} = r_{\text{child}} \omega$$

Given this, we can write the centripetal force required to keep the child riding on this disk as:

$$F_{\text{centripetal}} = \frac{m_{\text{child}} v_{\text{child}}^2}{r_{\text{child}}} = \frac{m_{\text{child}} (r_{\text{child}} \omega)^2}{r_{\text{child}}} = m_{\text{child}} r_{\text{child}} \omega^2$$

This force increases as the child moves from the center of the disk (where the force would be zero) to the edge. It increases dramatically, as the square of the angular velocity, when the disk spins more rapidly. And not surprisingly, the force required to keep traveling in a circle is larger for a more massive child.

But it is not only the child which requires a force to continue moving in a circle. Every part of the spinning disk requires such forces. The outermost ring of the disk is pulled inward by neighboring parts of the disk, which are themselves pulled inward by parts of the disk interior to them, all the way in to the center. As the disk spins faster and faster, the inner parts of the disk must pull inward with ever increasing force on the outer parts. Imagine the central ring of the disk. As it pulls inward on all the outer rings, they pull outward on it, stretching it and trying to pull it apart. If the disk spins too rapidly, these internal stresses may become too large, the material may break, and the whole thing may fly apart. This doesn't often happen on the playground of course, but the material stresses associated with rotation become very large in some of our technologies, things like large wind turbines and high speed centrifuges.



We know now that the sizes of these internal forces will increase as the square of the rate at which the object rotates, and also with the size and mass of the object rotating. In wind turbines, the rate of rotation is not extremely high, but the radius and mass of the rotating blades are very large. In an ultracentrifuge, the radius and mass are not so large, but the rate of rotation can be extreme, as high as 1200 rotations per second.

Getting an object rotating: torque and angular acceleration

The discussion above tells us something about the forces which must act inside an object which is spinning. How do we get an object spinning in the first place, and how might we predict the progress of its rotational motion? Recall how we did this with linear motion. We realized that if we knew the acceleration of every object, we could predict its future motion. Newton taught us to find the acceleration of an object by studying the forces which act on it. Force is related to acceleration by inertia. In rotational motion there is an exact parallel. We need to know angular accelerations, and so we study torques. Torques are related to angular acceleration by rotational inertia.

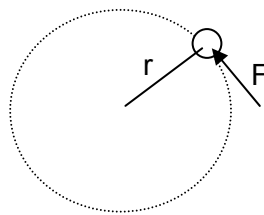
To do this for linear motion we use Newton's laws, in particular:

$$\vec{F} = \frac{d\vec{p}}{dt} = m\vec{a}$$

In rotational motion we would write:

$$\vec{\tau} = \frac{d\vec{L}}{dt} = I\vec{\alpha}$$

Where $\vec{\tau}$ is the torque, \vec{L} is the angular momentum, I is the rotational inertia (both to be defined in a moment), and $\vec{\alpha}$ is the angular acceleration.



Let's see how this comes about in a simple example. If I have a force acting on an object rotating in a circle like this, I will find that the force creates an acceleration

$$a_{\text{circle}} = \frac{F}{m}$$

along the circle. How does this linear acceleration relate to the angular acceleration of rotational motion? We saw above that, for an object traveling in a circle, linear velocity is related to angular velocity in a

simple way, and we can use this to derive the relation between linear acceleration and angular acceleration:

$$v_{\text{circle}} = r\omega \quad \text{and} \quad a_{\text{circle}} = \frac{dv_{\text{circle}}}{dt} = r \frac{d\omega}{dt} = r\alpha$$

so:

$$a_{\text{circle}} = r\alpha = \frac{F}{m} \quad \text{or} \quad F = mr\alpha$$

We know that torque is the quantity which generates rotation in the way that force generates motion. Recalling the definition of torque we can write:

$$\vec{\tau} = \vec{r} \times \vec{F} \quad \text{and} \quad \tau = r_{\perp} F = rF = r(mr\alpha)$$

$$\tau = mr^2\alpha$$

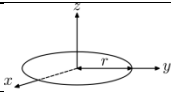
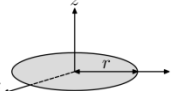
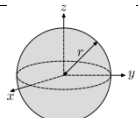
Note the similarity between this and Newton's second law for linear motion:

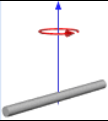
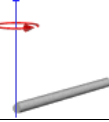
$$\vec{F} = m\vec{a} \qquad \vec{\tau} = I\vec{\alpha}$$

They are of just the same form, with only one variation. For linear motion the "inertia" of an object is determined just by its mass m . For rotational motion the "rotational inertia" of an object is determined both by its mass and the distance of the object from the center of rotation. We usually label this rotational inertia with the symbol I . If, instead of a single mass circling the center we have some more complex object (the blades of a wind turbine for example), the rotational inertia is determined by summing this product of mass times distance from the center squared:

$$I_{\text{total}} = \sum_i m_i r_i^2 \Rightarrow \int r^2 dm$$

These rotational inertias (which are sometimes called the "moment of inertia") can be calculated for many simple objects using the techniques of calculus, and values for a variety of specific shapes are given in the table below.

Description		Rotational inertia
Ring with mass m and radius r		mr^2
Solid disk with mass m and radius r		$\frac{1}{2}mr^2$
Solid sphere with mass m and radius r		$\frac{2}{5}mr^2$

Thin rod with mass m and length L rotating around its center		$\frac{1}{12}mL^2$
Thin rod with mass m and length L rotating around one end		$\frac{1}{3}mL^2$

Here is an example calculation for the rotational inertia of a uniform solid disk with mass m and radius r , rotating about its center. Break the ring into a series of thin concentric rings. Each contributes to the total rotational inertia an amount:

$$dI = m_{\text{ring}} r_{\text{ring}}^2 = \left[\left(\frac{m_{\text{disk}}}{\pi r_{\text{disk}}^2} \right) 2\pi r_{\text{ring}} dr \right] r_{\text{ring}}^2 = \frac{2m_{\text{disk}}}{r_{\text{disk}}^2} r_{\text{ring}}^3 dr$$

Summing this up over all the little rings from the center to the edge we get:

$$I_{\text{total}} = \int_0^{r_{\text{disk}}} dI = \int_0^{r_{\text{disk}}} \frac{2m_{\text{disk}}}{r_{\text{disk}}^2} r_{\text{ring}}^3 dr_{\text{ring}} = \frac{2m_{\text{disk}}}{r_{\text{disk}}^2} \int_0^{r_{\text{disk}}} r_{\text{ring}}^3 dr_{\text{ring}}$$

$$I_{\text{total}} = \frac{2m_{\text{disk}}}{r_{\text{disk}}^2} \frac{r_{\text{disk}}^4}{4} = \frac{1}{2} m_{\text{disk}} r_{\text{disk}}^2$$

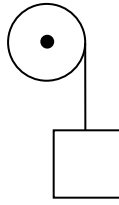
You may use some of these equations for rotational inertia in your homework, but it is important not to focus too much on specific cases. Concentrate instead on what this rotational inertia really is. Rotational inertia always depends on the mass of the object, but even more importantly on the distance of that mass from the center of rotation. It depends linearly on mass, but on the distance from the center squared. So if two objects have the same mass but one has mass farther from the center of rotation, it will have a larger rotational inertia.

You can see from the examples given above that rotational inertia is always the total mass of the object multiplied by some measure of the object's size squared and a unitless geometric factor which varies as the shape of the object changes.

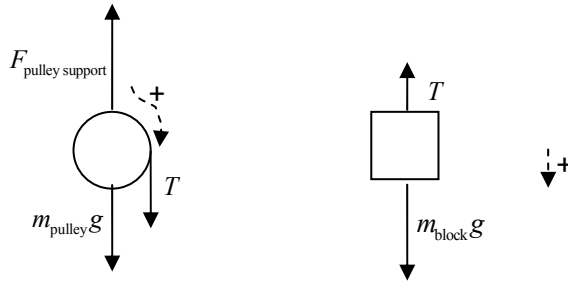
Look back now at the basic equation for rotational dynamics: $\tau = I\alpha$. It says that when a torque is applied to some object it will create an angular acceleration. The size of that angular acceleration will depend on the size of the torque, but also on the rotational inertia of the object; the larger the rotational inertia, the lower the angular acceleration.

A rotational dynamics example: a mass unwinding a rope from a pulley

Let's consider a simple example. Imagine a pulley with a rope wrapped around it, and a mass hanging from the rope.



The pulley has radius r_{pulley} and mass m_{pulley} , the block has mass m_{block} , and we will consider the rope to be massless. First, imagine what would happen if the pulley were massless. In this case, the block would accelerate downward with acceleration g , because the pulley would have no rotational inertia and would provide no resistance to motion. But what happens if the pulley has a nonzero mass? Start, as always in dynamics, with a free body diagram for each object, defining a positive direction of motion (or rotation) in each case.



Now for *each* of these bodies we know that:

$$\sum \vec{F} = m\vec{a} \quad \text{and} \quad \sum \vec{\tau} = I\vec{\alpha}$$

Consider the pulley first.

$$\sum F_y = F_{\text{pulley support}} - m_{\text{pulley}}g - T = 0 \quad \text{or} \quad F_{\text{pulley support}} = m_{\text{pulley}}g + T$$

So, whatever is holding up the pulley (presumably a bearing attached to the wall) must supply an upward force larger than the weight of the pulley alone. What about torques on the pulley? We know it will rotate about its center, so let's calculate the torques around that point.

$$\sum \tau_{\text{center}} = r_{\text{pulley}}T = I\alpha_{\text{pulley}} = \frac{1}{2}m_{\text{pulley}}r_{\text{pulley}}^2\alpha_{\text{pulley}}$$

This torque is positive because of the direction we chose to call positive rotation. The other forces generate no torque because they act at the center of rotation. Those forces, no matter what their size, will never make the object rotate.

OK, now let's do the same for the block:

$$\sum F_y = m_{\text{block}}g - T = m_{\text{block}}a_{\text{block}}$$

Where again, I have defined the downward direction, the direction the block will move, to be positive. There are no torques on the block, because neither of the forces which acts on it tends to make it rotate. OK, so we have two equations but three unknowns: T , α_{pulley} , and a_{block} . How to resolve this? We do it by noting that α_{pulley} is related to a_{block} . The faster the angular acceleration of the pulley, the faster the linear acceleration of the block. They are related by $a_{\text{block}} = r_{\text{pulley}}\alpha_{\text{pulley}}$. So:

$$T = \frac{I\alpha}{r} = \frac{\frac{1}{2}m_{\text{pulley}}r_{\text{pulley}}^2 a_{\text{block}}}{r_{\text{pulley}}^2} = \frac{1}{2}m_{\text{pulley}}a_{\text{block}}$$

And

$$m_{\text{block}}g - \frac{1}{2}m_{\text{pulley}}a_{\text{block}} = m_{\text{block}}a_{\text{block}} \quad \text{or} \quad \left(m_{\text{block}} + \frac{1}{2}m_{\text{pulley}}\right)a_{\text{block}} = m_{\text{block}}g$$

$$a_{\text{block}} = \left[\frac{m_{\text{block}}}{m_{\text{block}} + \frac{1}{2}m_{\text{pulley}}} \right] g$$

Notice what this means. One force is applied to accelerate the system: $m_{\text{block}}g$. But this force must not only accelerate the block, it must also set the pulley, which has some rotational inertia, rotating. So instead of $a_{\text{block}} = g$, we have the acceleration reduced by an amount which depends on the mass of the pulley. Does it depend on the radius of the pulley? No. Why not? Because although the rotational inertia of the pulley increases with radius, the torque available to rotate it increases too! So it doesn't matter if the pulley is big or small, just what its mass is.

What about the tension? We can find it now to be:

$$T = \frac{1}{2}m_{\text{pulley}}a_{\text{block}} = \frac{1}{2} \left[\frac{m_{\text{pulley}}m_{\text{block}}}{m_{\text{block}} + \frac{1}{2}m_{\text{pulley}}} \right] g$$

What are the limits of this:

- $m_{\text{block}} \gg m_{\text{pulley}}$: In this case the tension will approach zero, the acceleration of the block will approach the acceleration due to gravity, and the block will fall freely.
- $m_{\text{block}} = m_{\text{pulley}}$: In this case the tension will be one third the weight of the block, and the acceleration of the block will be two thirds of the acceleration due to gravity.
- $m_{\text{block}} \ll m_{\text{pulley}}$: In this case, the tension will be equal to the weight of the block, and the acceleration of the block will drop to zero. It will just hang there.

Rotational dynamics and rolling motion

Now let's consider another simple, but interesting example. If we release a ball with radius r on a slope, it begins to roll down it. In fact as it rolls down faster and faster, it undergoes both linear and angular acceleration. Since it undergoes linear acceleration down the slope, there must be a net force acting on it. Since it begins to rotate around its center, there must be a net torque acting on it.



How do we solve this problem? As always, we should draw a free body diagram then sum the forces and torques acting on the object. For this object, there is obviously a weight pulling it downward, and a normal force preventing it from passing through the slope beneath it. But there is another force too; a frictional force between the object and the surface beneath it. We know this frictional force must be there, because if it was not, the ball would simply slip down the slope, and would never begin to rotate. What kind of friction is it; static or kinetic? It could be either, but in many cases we will find that a rolling object will roll 'without slipping'. This is the normal way we see rolling motion work, and it should be quite familiar from the way wheels typically roll. When an object rolls without slipping, the point of contact is actually at rest relative to the surface beneath; it doesn't slip over the surface at all. Since there is no relative motion, this is static friction.

$$\begin{aligned}\sum F_{\text{along slope}} &= mg \sin(\theta) - F_{\text{static friction}} = ma_{\text{along slope}} \\ \sum F_{\perp \text{ to slope}} &= F_{\text{normal}} - mg \cos(\theta) = 0 \\ \sum \tau_{\text{around center}} &= rF_{\text{static friction}} = I\alpha\end{aligned}$$

If this ball rolls without slipping we can relate the angular and linear acceleration just as we did above: $a = r\alpha$. This allows us to solve for the size of the static friction in terms of the acceleration:

$$rF_{\text{static friction}} = \frac{Ia}{r} \quad \text{or} \quad F_{\text{static friction}} = \frac{Ia}{r^2}$$

Returning to the first equation

$$mg \sin(\theta) - \frac{Ia}{r^2} = ma$$

$$a = \left[\frac{g \sin(\theta)}{\left(1 + \frac{I}{mr^2}\right)} \right]$$

If our ball rolling down the slope is a solid sphere, we can look up its rotational inertia and find $I = \frac{2}{5} mr^2$. As a result, the acceleration down the slope will be:

$$a = \frac{5}{7} g \sin(\theta)$$

The acceleration of the ball down the slope is reduced from what it would be with no friction. Looking back at the equations above, we can see what would happen if there was no friction. In this case, we would find:

$$a_{\text{no friction}} = g \sin(\theta)$$

The ball would simply slip down the slope. There would be no torque, and it would never begin to rotate, it would just accelerate along the slope.

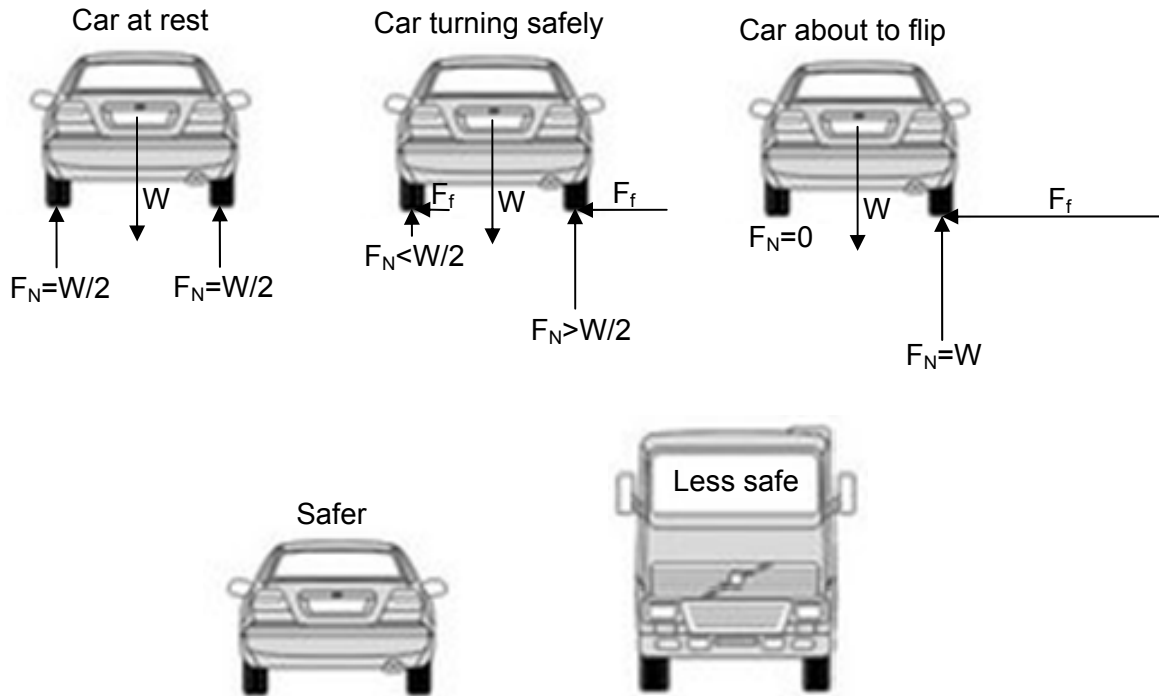
The general relation for the acceleration derived above tells us still more. It says that if the rotational inertia of the object rolling down the slope is bigger, its acceleration will be smaller. So if I take the hollow sphere and solid sphere and rolled them both down a slope, the solid sphere, which has more of its mass near the center of rotation, will win the race to the bottom.

There is another interesting thing to notice here. Since the rotational inertia I will always be proportional to the total mass of the object m , the ratio I/mr^2 will always be independent of m . The size of this rolling acceleration depends only on the distribution of the mass in the object, not on its total mass.

Rotational dynamics and roll-over accidents

Another issue central to safely turning corners comes from the fact that the frictional force between your tires and the ground is applied at the ground, right at the point where the tires touch the road. When conditions are wrong, this can cause the car to begin to flip, creating one of the famous roll-over accidents which so plague sport-utility vehicles. The pictures below illustrate something of how this works. The first shows a car at rest, the second a car turning a corner gradually (and safely), and the third turning a corner so rapidly that it is about to flip. Once the car begins turning the corner, the sum of the forces on it is no longer zero; it actually is accelerating toward the center of the circle. For a car turning on level ground, the force which creates this centripetal acceleration is friction with the ground; usually static friction.

Something else begins to happen as well. During the turn, a larger fraction of the car's weight is supported on the outside wheels than the inside wheels; the load shifts. This load shift is seen in all accelerating vehicles, even those accelerating straight forward. As the acceleration becomes larger, the load shift increases, until eventually all the weight is supported on the outside wheels for a turning car, or back wheels for an accelerating one. If the car turns still more suddenly, it will tip.



While this is an accurate representation of what happens, it doesn't explain it at all. **Why** does the load shift? What are the details of this shift, and how does it lead to the final balance condition just before the car tips?

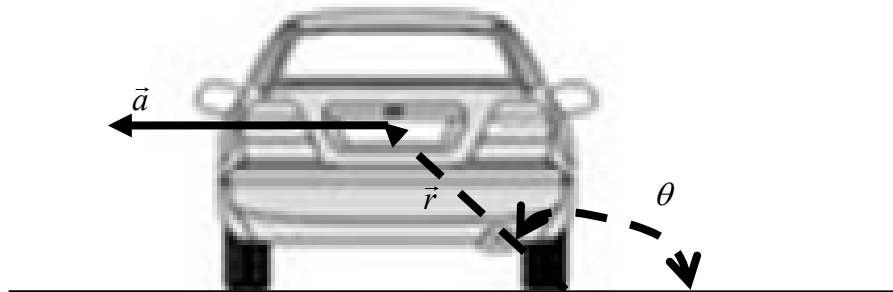
To understand this, consider the car at rest first, and then when it is turning. We will analyze it in an inertial frame connected to the road, not an accelerating reference frame connected to the car. We will sum torques around the point on the road where the right wheel touches. We will call the wheel base b and the height of the car h . For the car at rest, we write:

$$\begin{aligned} \sum F_x &= 0 \\ \sum F_y &= F_{NL} + F_{NR} - mg = 0 \\ \sum_{\text{right wheel}} \tau &= mg \frac{b}{2} - F_{NL} b = 0 \\ F_{NL} &= \frac{mg}{2} \quad \text{and} \quad F_{NR} = \frac{mg}{2} \end{aligned}$$

No surprise here. This is the simple case where the weight of the car is equally shared between the tires on the left and right.

Next, consider the general case when the car is turning. Now the car does experience acceleration to the left; the centripetal acceleration. It is caused by the frictional force exerted on the car by the road. The size of the centripetal acceleration depends on the cars speed and the radius of the circle along which it is turning: $a_c = v^2/r$.

More important, and less obvious, the car now experiences an angular acceleration about the fixed point on the road! To see this, consider the following picture:



Because the center of mass of the car is actually accelerating to the left with a linear acceleration $a = v^2/r$, the angular acceleration α is non-zero. We can work out what it is in terms of the linear acceleration with a little algebra:

$$\begin{aligned}\cos(\theta) &= \frac{x}{y} \\ \frac{d^2x}{dt^2} &= -\frac{v^2}{r} \\ \frac{d^2 \cos(\theta)}{dt^2} &= -\cos(\theta) \frac{d^2\theta}{dt^2} = \frac{1}{y} \frac{d^2x}{dt^2} \\ -\frac{x}{y} \frac{d^2\theta}{dt^2} &= -\frac{1}{y} \frac{v^2}{r} \\ \alpha = \frac{d^2\theta}{dt^2} &= \frac{v^2}{rx} = \frac{2v^2}{rb}\end{aligned}$$

As soon as there is linear acceleration there really is angular acceleration about a fixed point on the ground. Once we acknowledge this, we can write the sum of torques and forces again as:

$$\begin{aligned}\sum F_x &= -F_{fL} - F_{fR} = ma_x = -\frac{mv^2}{r} \\ \sum F_y &= F_{NL} + F_{NR} - mg = 0 \\ \sum_{\text{right wheel}} \tau &= mg \frac{b}{2} - F_{NL} b = I\alpha = I \frac{2v^2}{rb}\end{aligned}$$

The rotational inertia of a rectangle of base b and height h around its corner is given by:

$$I_{\text{rectangle about corner}} = \frac{m(b^2 + h^2)}{12} + m(b^2 + h^2) = \frac{13m(b^2 + h^2)}{12}$$

But let's leave this as I , since the car surely has a somewhat different form than a rectangle. Whatever it is, it will be proportional to mass and size². These equations allow us to find the normal force on the left and right:

$$F_{NL} = \frac{mg}{2} - I \frac{2v^2}{rb^2} \quad \text{and} \quad F_{NR} = \frac{mg}{2} + I \frac{2v^2}{rb^2}$$

These equations explain why there must be a load shift from evenly balanced to all on the left. They also tell us that the crisis will occur when:

$$F_{NL} = 0 \quad \text{or} \quad \frac{v^2}{r} = \frac{mg}{4} \frac{b^2}{I}$$

If you write this down for a rectangle, you get:

$$\frac{v^2}{r} = \frac{mg}{4} \frac{12b^2}{13m(b^2 + h^2)} = g \frac{3b^2}{13(b^2 + h^2)} = g \frac{3}{13 \left(1 + \frac{h^2}{b^2}\right)}$$

The limiting centripetal acceleration is just some multiple of g . For a square it would be $3g/26$, or about $g/8$. For a rectangle twice as high as wide, this would be $3g/78$, or about $g/25$. For a rectangle half as high as wide, this would be about $g/5$.

To allow a turning acceleration as large as g , you would have to have a car with a rotational inertia about that axis given by:

$$I \geq \frac{mb^2}{4}$$

So there is non-zero torque and angular acceleration around a fixed point at the location of the right tire. The size of that angular acceleration increases as the centripetal acceleration of the car increases. If this is to be true, there must be a load shift, and the car will tip just after all the load has slipped to the outside tire.

When a turning object is tall and thin, like a bicycle or a person, this problem is severe indeed. In this case it is usually handled by tilting inward while turning so that the sum of the normal force and frictional force applied when the foot (or tire) meets the ground passes straight through the center of mass of the object. When the required centripetal force coming from friction is small compared to the normal force supporting the weight, the tilt is small. As it increases, the tilt becomes more dramatic, as illustrated in the images of turning horse, water skiers, cheetahs, and bicyclists. When the required centripetal acceleration equals the weight, you should expect this tilt to be 45° .



Introduction to angular momentum

We found that one of the linchpins of linear mechanics was Newton's first law, the idea that any object will continue in a state of uniform motion unless some force acts upon. This idea, which was well developed by Galileo's time, is the key to dynamics, because it focuses our attention on change. It says that when we see change, we know there is a force, and that if we wish to understand change, we need to understand forces.

There is a very similar (and in fact sort of more apparent) principle at play in rotational motion. When an object begins to rotate, it will continue to rotate unless some force acts to end the rotation. But it can't be just any force, it has to be a force capable of exerting a torque. There is a kind of angular momentum which exists independent of linear momentum. What we want to do now is quantify this and see where it fits into our parallel between linear and angular motion.

First of all, the idea of linear momentum is expressed in:

$$\vec{p} = m\vec{v}$$

mass times velocity. Our parallel with rotational motion leads us to suspect that the angular momentum would be:

$$\vec{L} = I\vec{\omega}$$

where the symbol \vec{L} represents the angular momentum, I the rotational inertia, and $\vec{\omega}$ the angular velocity. For this discussion, we're really going to stick to a vector expression for both angular momentum and angular velocity (just as we do for momentum and velocity). You will see why this is important in a bit.

There is another, also instructive way to write the angular momentum. Just as the torque is:

$$\vec{\tau} = \vec{r} \times \vec{F}$$

The angular momentum can be written:

$$\vec{L} = \vec{r} \times \vec{p}$$

Is this the same as what we wrote before? Imagine a little ball of mass m spinning in a circle of radius r . What is its angular momentum?

$$\text{First definition: } L = I\omega = (mr^2)\omega$$

$$\text{Second definition: } L = r \times p = r(mv) = mr^2\omega$$

so these two quantities are the same.

Conservation of angular momentum

How does angular momentum change? Remember that:

$$\vec{F} = \frac{d\vec{p}}{dt}$$

The force is equal to the time rate of change of the momentum. Now we might expect in a parallel way that the torque is equal to the time rate of change of the angular momentum. Is this true?

$$\vec{\tau} = I\vec{\alpha} = I \frac{d\vec{\omega}}{dt} = \frac{d(I\vec{\omega})}{dt} = \frac{d\vec{L}}{dt}$$

What does this mean? *Most* important, if there are no torques, the angular momentum does not change, it is conserved. This is just what we set out to show. Now let's think about what it means.

How could L change? Linear momentum is determined by $\vec{p} = m\vec{v}$. An object usually cannot change its mass, so linear momentum usually changes by changing the velocity: $d\vec{p} = m d\vec{v}$. Angular momentum $\vec{L} = I\vec{\omega}$. Unlike the linear momentum, L can change either by changing the angular velocity *or* by changing the moment of inertia. The angular momentum in an isolated object can be conserved while having both I and ω change. How does this work? If no torques act we can be sure that the angular momentum doesn't change, so:

$$L_{\text{final}} = L_{\text{initial}} \quad \text{or} \quad I_{\text{final}}\omega_{\text{final}} = I_{\text{initial}}\omega_{\text{initial}} \quad \text{or} \quad \frac{\omega_{\text{final}}}{\omega_{\text{initial}}} = \frac{I_{\text{initial}}}{I_{\text{final}}}$$

How can an object change its rotational inertia I ? All it has to do is rearrange the locations of the masses which make it up.

If you have ever watched gymnasts, divers, or the Winter Olympics you have seen this happen a lot. When a figure skater begins to spin and then pulls her arms in, she spins faster and faster. She does this not because she is applying a torque to make herself spin faster, but just because of the conservation of angular momentum. A diver leaves the board rotating slowly, then pulls himself into a ball spinning very rapidly, and at the end of the dive again extends his body, slowing his rotation so that he might enter the water smoothly.

There is a fascinating and very important consequence of angular momentum conservation. If we set a rigid body spinning, and make sure no torques can act on it, its angular momentum will be conserved. In particular, this means that the direction of its angular momentum vector will never change. This fact allows us to set an object spinning with its axis pointing, say, North. Then we can lock it up in a box run back and forth, go inside a building, and still know, from the direction it's spinning, exactly which way is North. This fact is used extremely generally in all sorts of guidance systems.

There is a very important point to note about all this. It may seem that angular momentum is not really new, that somehow it's just an obvious extension of linear momentum, something which must be guaranteed by Newton's first law. This is not the case, in part for the following reason. Angular momentum is often conserved in situations in which the linear momentum is not. This is because very often in nature forces act which, although they change the linear momentum of an object, do not affect its angular momentum. We can see this by thinking again about skating. When a figure skater pulls her arms in as she spins, she has to apply an inward force to pull them in. This force alters the linear momentum of her arms, but because the force is central, because it passes directly through the center of rotation, it generates no torque, and hence doesn't change the angular momentum.

Any force which can act only along the line directly between two objects is called a "central force". Good examples of central forces include gravity, or the tension in a rope; each can only pull objects together directly along the line between them. Because of this, neither force can ever generate a torque, neither can ever change the angular momentum of a system.

Think about a planet orbiting the sun. At every point the only force on it acts directly between the center of the planet and the center of the sun. Such a force can never exert a torque on the planet, and the angular momentum of the planet as it orbits around the sun can never change. It is conserved.

Torque, angular momentum change, and precession

To get a clear understanding of how angular momentum does change, it's useful to take a step back and think first about how linear momentum changes. We know that

$$\vec{F} = \frac{d\vec{p}}{dt}$$

So if $\vec{F} \parallel \vec{p}$, the force changes only the magnitude of the momentum. This is the case we examined when we talked about getting started in linear motion, forces that act only along the direction of motion. If, on the other hand, $\vec{F} \perp \vec{p}$, the force changes only the direction of the momentum, and not its magnitude. This is what happens in, for example, uniform circular motion.

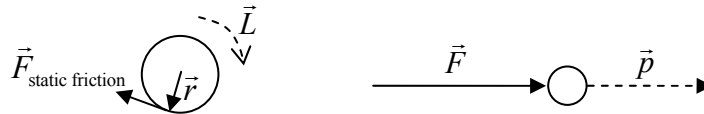
In a very parallel way:

$$\vec{\tau} = \frac{d\vec{L}}{dt}$$

So when $\vec{\tau} \parallel \vec{L}$, the torque changes only the magnitude of the angular momentum. This is what happens in simple cases of angular acceleration, as in what happens when a ball rolls down a slope. Something more surprising occurs in the less common case when $\vec{\tau} \perp \vec{L}$. In this case, the torque changes only the direction of the angular momentum, and not its magnitude.

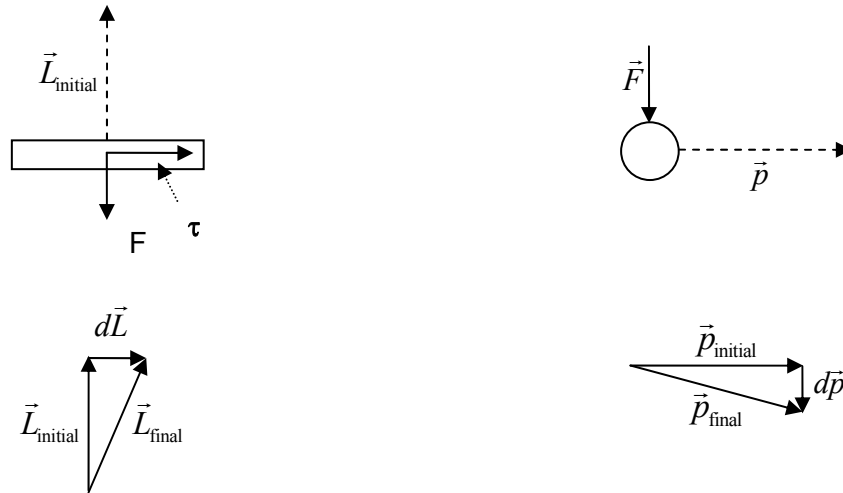
Let's look at examples of each:

1. First $\vec{\tau} \parallel \vec{L}$: This is what happens when a ball rolls down a slope. Here we have a torque generated on the ball which, by the right hand rule, is into the board. As the ball begins to spin its angular momentum is also into the board. So the torque and the angular momentum are parallel, and what happens is ordinary angular acceleration. This is just like the case where the linear force is parallel to the momentum.



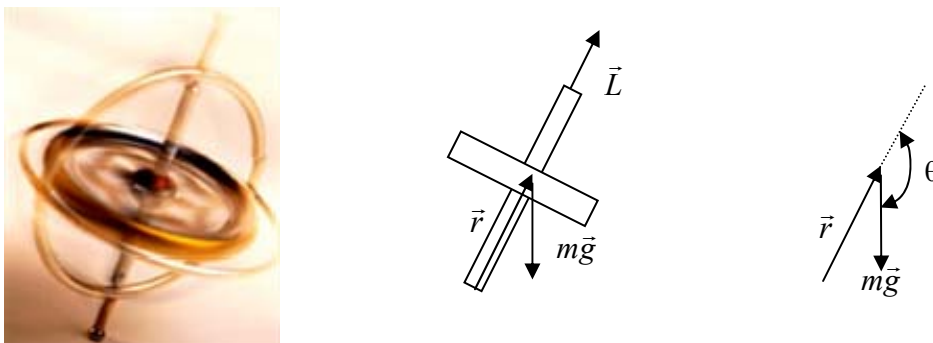
2. Second, $\vec{\tau} \perp \vec{L}$. This is a little trickier to see. Imagine watching a disk spinning so that its angular momentum is up. You look at it from the side. If we push down on the edge of the disk which is closest, what will happen? We will create a torque around the center of mass. What direction will this torque be in? The radius vector \vec{r} is pointing towards us, and the force vector \vec{F} is pointing down. So if you point the fingers of your right hand out of the page, and curl them down, your thumb will go to the right. So the torque vector is to the right:

it is perpendicular to the angular momentum vector. What does this mean for the way the angular momentum changes? We can write $d\vec{L} = \vec{\tau} dt$, so just as a force perpendicular to the linear momentum changes only the direction of the linear momentum, so too a torque perpendicular to the angular momentum changes only the direction of the angular momentum.



To sum up: if we apply a torque along the same axis as the angular momentum, we get angular acceleration, a change in the magnitude of the angular momentum. If we apply a torque perpendicular to the direction of the angular momentum we get a change in the direction of the angular momentum, but no change in its magnitude.

A simple example which you may have seen is a tilted gyroscope. What happens here is outlined in the picture below:



The angular momentum through the support point at the bottom of the gyroscope is shown as \vec{L} . The weight of the object acts at the gyroscope's center of mass. This means that the weight applies a torque around the support point:

$$\vec{\tau} = \vec{r} \times \vec{F} = rm\vec{g} \sin(\theta) \text{ (into the page)}$$

This will cause changes in the angular momentum into the page, perpendicular to its current direction. As time goes by, at each instant the change in angular momentum will remain perpendicular to the angular momentum. The gyroscope will "precess", with its angular momentum vector traveling around in a circle in much the same way that the linear momentum vector of an object in uniform circular motion travels in a circle. The rate of precession depends on the magnitude of the torque and the magnitude of the angular momentum.

Here is an example of how we might calculate this precession rate. The angular momentum will change in a way which is determined by the torque: $d\vec{L} = \vec{\tau} dt$. So during a little time period dt the angular momentum \vec{L} will turn through an angle:

$$d\theta = \frac{d\vec{L}}{\vec{L}} = \frac{\vec{\tau} dt}{\vec{L}}$$

It will spin all the way around when it turns through an angle 2π , or when:

$$2\pi = \frac{\tau \Delta t}{L} \quad \text{or} \quad \Delta t = \text{period of precession} = \frac{2\pi L}{\tau}$$

In other words, decreasing L leads to more rapid precession, increasing L to less. A rapidly spinning gyroscope should precess slowly, while a slowly spinning gyroscope should precess quickly. Increasing τ leads to more rapid rotation, decreasing it to less.

Rotational motion is relatively rare in living systems, but plays a major role in our technologies. For this reason, it is useful to understand its essentials. Having explored it a bit, we will set it aside now and move on to a second example of 2D motion which is simple to analyze and useful for modeling more complex motions.

8.4 A second useful model: motion with a single constant force

Another important example of two dimensional motion occurs when an object feels a single force constant in both magnitude and direction. When this happens, the object moves in a parabolic path, with the curvature of the parabola in the direction of the single force which acts. Having a single, perfectly constant force act is not especially common. But there are times when this is approximately true, periods of time when the total force on an object is nearly constant. During these times, the motion we describe in detail here can provide a useful model, just as uniform circular motion can provide a useful model for the acceleration of an object traveling with approximately constant speed on a curved path.

The most famous example of this kind of motion is an object thrown or fired near the Earth's surface. Such an object receives all the force which gets it going at the beginning, then flies without anything pushing it forward. It moves in what is called 'ballistic' motion. In most real-world cases, such an object will experience not one, but two forces; the downward force of gravity and an air friction force resisting its motion. We will consider this full, more complicated motion in a bit. But we will begin with a simpler model, one in which we imagine we can ignore air friction. An object in ballistic motion without air

friction feels a constant downward force, but no horizontal force at all. What would the motion of such an object be like?

We know how that motion would *change* already. The constant downward force would create a constant downward acceleration. This acceleration would continually change the vertical component of the velocity. Whatever vertical velocity it begins with will continuously become more and more negative. What about the horizontal velocity? We have assumed that there is no force in the horizontal direction, so there will be no horizontal acceleration. As a result, whatever horizontal motion the object has to start will stay exactly the same throughout the motion.

This problem of ballistic motion has many, many practical applications. The term ballistic is a modernization of the Latin 'ballista', the name applied to a kind of giant crossbow used as a siege engine. For reasons of military history, ballistic motion is often called projectile motion. Before we work out the details, let's look at a little physics history.

Projectile motion and superposition in physics

One of the reasons for resolving a vector into its components is that the motion of objects along different directions can be independent. This allows us to consider each component of the total motion independent of the others. The oldest application of this idea in physics, first clearly analyzed by Galileo, is to projectile motion. The motion of a projectile includes several parts. The object rises and falls, but also travels horizontally while doing this. This much (and more) was well known to lots of people before Galileo.

The crucial contribution Galileo made to our understanding of how to describe this motion is contained in the following quotation from his book 'Two New Sciences'. In it he speaks of 'naturally accelerated' motion. By this he means motion accelerated by gravity.

I propose to set forth those properties which belong to a body whose motion is compounded of two other motions, namely one uniform and one naturally accelerated....This is the kind of motion seen in a projectile....Imagine any particle projected along a horizontal plane without friction...This particle will move along this plane with a motion that is uniform and perpetual, providing the plane has no limits. But if the plane is limited and elevated, then the moving particle...will, on passing over the edge of the plane, acquire, in addition to its previous uniform and perpetual motion, a downward propensity due to its own weight; so that the resulting motion... is **compounded of one which is uniform and horizontal, and one which is vertical and naturally accelerated.**

Galileo is making a quite surprising assertion about the physics of motion here. He is saying that an object's horizontal motion has no effect on how it falls, and further, that its falling has no effect on its horizontal motion. He says that the motion is a **superposition** of two independent motions. This idea of superposition is a very important one in physics, one we will return to again and again.

This is not a statement about vectors or math. It's a statement about physics, about how the world really works. It cannot be proven by deduction, but only by experiment. Also note carefully that there are two statements here:

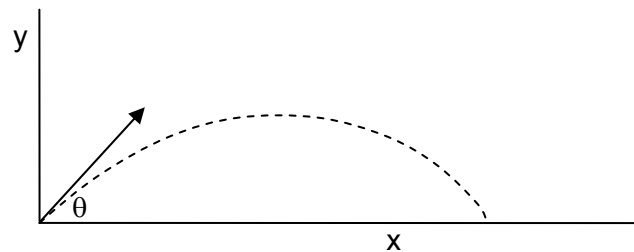
1. A horizontal velocity does not affect the way an object falls
2. The fact that an object is falling does not affect its horizontal velocity

They are not the same, and Galileo experimentally tested each. Each time, he had to work to reduce the complicating effects of air friction, then employ reason to imagine what would happen if he could completely eliminate it. He confirmed the first principle by showing that two balls, one moving horizontally and one not, fell at the same rate. He confirmed the second by showing that two balls, one moving vertically and one not, continued in the same horizontal motion. These experiments supported the principle of superposition in this case, and helped to introduce this important idea into physics.

Once Galileo recognized that he could separate the horizontal and vertical motions in this way, he was able to work out in detail the path of such a friction-free projectile, as indeed we are about to do.

Frictionless projectile Motion: the details

Let's analyze this motion in detail. Imagine an object which starts out with some initial velocity \vec{v}_i , which



has a magnitude v_i and a direction an angle θ above the horizontal:

Notice that the picture above is not one of our usual position-time graphs. It shows instead two spatial dimensions, horizontal and vertical, with the initial velocity v_i drawn as a vector. The path this projectile will ultimately follow is suggested by the dashed line. We can write this initial velocity vector in terms of horizontal and vertical components as:

$$\vec{v}_i = v_{ix}\hat{x} + v_{iy}\hat{y} = v_i \cos(\theta)\hat{x} + v_i \sin(\theta)\hat{y}$$

How does the motion progress after the launch? As Galileo realized, the two motions, horizontal and vertical should be completely independent, so we really have two separate motions to describe:

Horizontal: The horizontal motion experiences no acceleration, so it continues at constant horizontal speed, and the distance traveled is described by:

$$v_{fx} = v_{ix} = v_i \cos(\theta)$$

$$x_f = x_i + v_i \cos(\theta) \Delta t$$

Vertical: The vertical motion experiences just the free falling acceleration due to gravity that we have been analyzing so far in class, so its motion is described by:

$$v_{fy} = v_{iy} + a\Delta t = v_i \sin(\theta) - g\Delta t$$

$$y_f = y_i + v_{iy}\Delta t + \frac{1}{2}a(\Delta t)^2 = y_i + v_i \sin(\theta)\Delta t - \frac{1}{2}g(\Delta t)^2$$

This vertical motion is exactly what we analyzed in the previous chapter when we talked about an object thrown straight upward with an initial speed v_{iy} .

We now have equations for x_f and y_f as a function of time. We can combine these, eliminating the time variable Δt from the two equations, to determine the path $y(x)$. Noting that $\Delta x = x_f - x_i$ and $\Delta y = y_f - y_i$, we can write:

$$\Delta t = \frac{\Delta x}{v_i \cos(\theta)}$$

$$\Delta y = v_i \sin(\theta) \frac{\Delta x}{v_i \cos(\theta)} - \frac{1}{2}g \left(\frac{\Delta x}{v_i \cos(\theta)} \right)^2$$

$$\Delta y = \tan(\theta)\Delta x - \left(\frac{g}{2v_i^2 \sin^2(\theta)} \right) (\Delta x)^2$$

Notice what this says. The distance traveled in the y direction (Δy) is a quadratic function of the distance traveled in the x direction (Δx). This tells us that the motion is parabolic: the shape of the path is a portion of a parabola. That's it really; we now know how "projectiles" move.

As always, we have to remember precisely what we solved for here. These equations describe the motion of an object moving under the influence of a single constant force, in this case the weight. This will be a reasonable model for the motion of an object through the air only when conditions are such that the force of air friction is *always* small compared to the object's weight. As we have argued before, this will happen only when the object is relatively heavy and moving relatively slowly. When this is the case, these equations will provide a reasonably accurate model for the motion, which we can describe as follows:

$$\text{Position: } \vec{s} = \Delta x \hat{x} + \Delta y \hat{y} = (x_i + v_i \cos(\theta)\Delta t) \hat{x} + \left(y_i + v_i \sin(\theta)\Delta t - \frac{1}{2}g(\Delta t)^2 \right) \hat{y}$$

$$\text{Velocity: } \vec{v} = v_x \hat{x} + v_y \hat{y} = v_i \cos(\theta) \hat{x} + (v_i \sin(\theta) - g\Delta t) \hat{y}$$

$$\text{Acceleration: } \vec{a} = a_x \hat{x} + a_y \hat{y} = -g \hat{y}$$

Several examples of frictionless projectile motion

Armed with this model for projectile motion, we can answer many possible questions about real projectiles. Let's do a simple example:

Imagine that we launch a baseball from the ground with an initial velocity $v_i = 20 \text{ m/s}$ at an angle $\theta = 30^\circ$ above the horizontal, how far does it travel before returning to the ground?

Start by considering the vertical motion alone. It has an initial vertical velocity of

$$v_{iy} = v_i \sin(\theta) = (20 \text{ m/s}) \sin(30^\circ) = 10 \text{ m/s}$$

How long does it stay in the air? This is exactly like the problem we did earlier, in which we asked how long does a ball stay in the air when we throw it straight up with an initial speed of 10 m/s. In that case, we used the fact that it will return to the ground traveling at the same speed it left to find the time Δt_{flight} :

$$v_{fy} = -v_{iy} = v_{iy} - g\Delta t_{\text{flight}} \quad \text{or} \quad \Delta t_{\text{flight}} = \frac{2v_{iy}}{g} = \frac{2v_i \sin(\theta)}{g}$$
$$\Delta t_{\text{flight}} = \frac{20 \text{ m/s}}{9.8 \text{ m/s}^2} = 2.04 \text{ s}$$

OK, so the projectile is in the air for about 2.04 s, how far does it travel in this time? The horizontal motion is governed by:

$$\Delta x = v_{ix} \Delta t_{\text{flight}} = v_i \cos(\theta) \Delta t_{\text{flight}}$$
$$\Delta x_{\text{range}} = (20 \text{ m/s}) \cos(30^\circ) (2.04 \text{ s}) = 35.3 \text{ m}$$

Is this a reasonable answer? The throw is pretty hard, an initial speed of nearly 50 miles per hour, and travels a pretty good distance, about 115 feet, about the distance from second base to home plate on a baseball diamond. Just as an aside, this initial velocity is pretty high, so friction would probably not be negligible in this case. We will check on this more carefully in the next section.

Consider what we have done here; realizing that the vertical motion (in this case) determines the duration of the flight, we were able to determine the horizontal distance traveled. Here we see some of the elements of typical projectile problems: because the motions are coupled, we can usually use one of the motions to constrain the other. You ought by now to be able to see how we would find out how high the ball rises during this flight too.

Looking back at the relation for horizontal position Δx given above, and at the relation for Δt_{flight} derived from the vertical motion, we can write an interesting new relation for the total horizontal distance traveled (the range of the ball):

$$\Delta x_{\text{range}} = v_i \cos(\theta) \Delta t_{\text{flight}} = v_i \cos(\theta) \left(\frac{2v_i \sin(\theta)}{g} \right) = 2 \sin(\theta) \cos(\theta) \frac{v_i^2}{g}$$

There is a nice way to rewrite this using the trigonometric identity:

$$2 \sin(\theta) \cos(\theta) = \sin(2\theta)$$

To obtain the famous "range equation":

$$\Delta x_{range} = \left(\frac{v_i^2}{g} \right) \sin(2\theta)$$

What does this expression tell us? If we launch an object on level ground with some initial speed v_i , at some angle θ above the horizontal, and if we can ignore air friction, this relation tells us how far it will go before it strikes the ground.

What is a general relation like this good for? Of course we can use it to calculate the specific range for any particular case. But more importantly, it tells us how the result (the range) depends upon the input variables. From this general form we can see that if we double the initial speed, we quadruple the range. We also see that the maximum range occurs when the launch angle is 45° , because the maximum of $\sin 2\theta$ occurs at 45° . Galileo arrived at this same result, though his argument was couched in more explicitly geometric terms.

A caution: this range equation, like so many others throughout this class, applies *ONLY* when the conditions for which it was derived are satisfied. This requires that the *only* force acting is the downward force due of gravity. If air friction is important for your projectile, the range predictions provided by this model will be quite inaccurate. If, however, the forces due to air friction are always small compared to the downward force of gravity, this range equation should be OK. Again, Galileo acknowledged the same, and addressed the question directly:

‘No firm science can be given of such events of heaviness, speed, and shape, which are variable in infinitely many ways. Hence to deal with such matters scientifically, it is necessary to abstract from them. We must find and demonstrate conclusions abstracted from the impediments, in order to make use of them in practice under those limitations that experience will teach us’

And

‘in projectiles that we find practicable, which are those of heavy material and spherical shape, and even in [others] of less heavy material, and cylindrical shape, as are arrows, launched [respectively] by slings or bows, the deviations from exact parabolic path will be quite insensible’

He did not, indeed could not, attempt to solve the more complex, real world problem of ballistic motion *with* air friction. That required centuries of further work, including Newton’s discovery of the laws of motion, and a more precise understanding of fluid friction achieved only in the late 19th century.

There are other detailed assumptions here. Our derivation assumes a launch from level ground at an angle θ above the horizontal. This range equation does not apply if (for example) you are firing onto, or off of, or at, a wall or a cliff. For this reason, I strongly urge you to always begin from the basic assumption of independent horizontal and vertical motion and use what you know about the problem to derive the solution.

Consider a second example to see how this works.

The large rainforest tree *Tetraberlinia Moreliana*, like many canopy trees, spreads its seeds by ‘explosive dispersal’¹. The casings of its seed pods develop enormous tension as they dry, then burst suddenly, blasting the seeds out away from the tree with remarkable velocities. Seeds are launched from the top of the tree, perhaps 35 m above the ground, typically at angles of around 15° above the horizontal. Amazingly, they are sometime found more than 60 m from the tree. While air friction is undoubtedly important here, we can use our simple friction free model to get some idea of the velocities with which these seeds are launched.

How might we find this velocity? In this case, we will need to use both vertical and horizontal motion combined.

$$y_f - y_i = -35 \text{ m} = v_{iy} \Delta t - \frac{1}{2} g (\Delta t)^2$$

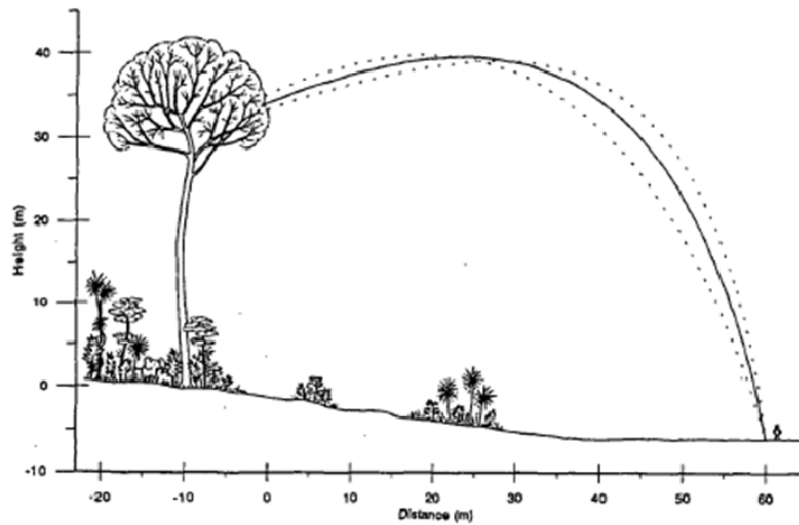
$$x_f - x_i = 60 \text{ m} = v_{ix} \Delta t$$

$$v_{iy} = v_{ix} \tan(\theta) = v_{ix} \tan(15^\circ)$$

These three equations with three unknowns (v_{ix} , v_{iy} , and Δt) can be solved to find the magnitude of the initial velocity:

$$v_i = \sqrt{v_{ix}^2 + v_{iy}^2} = 19.25 \text{ m/s}$$

This is a very substantial velocity, more than 40 miles per hour. Even if there were absolutely no effect of air friction, this tree would have to launch its seeds at this very high speed to spread them 60 m from its base. But of course there is air friction. Will it be important, and how would it affect the required launch velocity?



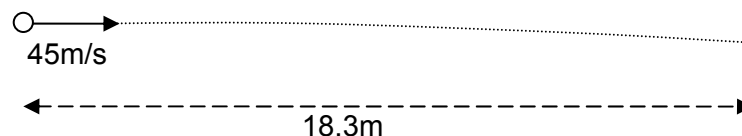
Air friction will be important when the frictional forces become comparable the weight of the projectile. In this case the seeds are fairly small; little disks with diameters of about 3 cm and masses of about 2.5 grams. Since they have small weights, and must be moving rather fast, we would expect friction to play a very important role indeed. How will it affect the motion? Since friction always resists the motion, we can be sure that the seeds are actually launched with speed *larger* than the simplest case values we calculated here. Calculations which include air friction suggest real launch velocities of nearly 40 m/s, or about 90 miles per hour. The initial explosion of these seed pods is said to sound like a small caliber hand-gun. Best not to mess with them when they're ready to burst!

Knowledge of projectile motion can also help us to better appreciate many different athletic endeavors. An analysis using these tools suggests why baseball is such a difficult sport.

A pitcher may throw at 100 miles per hour, and the ball must travel 60 feet to the batter

$$100 \text{ mph} * (1 \text{ m/s} / 2.24 \text{ mph}) = 45 \text{ m/s}$$

$$60 \text{ ft} * (1 \text{ m} / 3.28 \text{ ft}) = 18.3 \text{ m}$$



How long does it take to travel this distance?

$$\Delta x = 18.3 \text{ m} = v_{ix} \Delta t \quad \text{or} \quad \Delta t = \frac{18.3 \text{ m}}{45 \text{ m/s}} = 0.41 \text{ s}$$

How far does it drop in this time?

$$\Delta y = v_{iy} \Delta t - \frac{1}{2} g (\Delta t)^2 = -\frac{1}{2} (9.8 \text{ m/s}^2) (0.41 \text{ s})^2 = -0.82 \text{ m} \approx -32''$$

When the ball is half-way, it has dropped only one fourth as much:

$$\Delta y_{\frac{1}{2}} = -\frac{1}{2} g \left(\Delta t_{\frac{1}{2}} \right)^2 = -\frac{1}{2} (9.8 \text{ m/s}^2) (0.2 \text{ s})^2 = -0.20 \text{ m} \approx -8''$$

Now think about what happens if the initial speed is reduced to 85mph (37.9m/s):

$$\Delta t = \frac{18.3 \text{ m}}{37.9 \text{ m/s}} = 0.48 \text{ s (only 15% different)}$$

$$\Delta y = 1.1 \text{ m} (\approx 44'')$$

$$\Delta y_{\frac{1}{2}} = 0.28 \text{ m} (\approx 11'')$$

The batter, trying to predict where the ball will be when it arrives, has to look out and see how much the ball drops at the half way point. The difference he's looking for is just a few inches, while the ball is still about 10 meters away. Then in the remaining 0.2 seconds, the batter must adjust their swing for a pitch which may differ in location by 12", all the while swinging a bat which is officially "not more than 2 3/4" in diameter at the thickest part". With these calculations in hand, it is perhaps not so surprising that someone who can reliably succeed in this task even one third of the time might be so highly valued in the major leagues.

8.5 More complex cases: projectiles with air friction and beyond

In this chapter we first explored the basics of the connection between the two and three dimensional paths of objects and the forces which act on them. Then we explored the details of two special cases of 2D motion; uniform circular motion, and motion under the influence of a single constant force. In each of these two cases, we were able to derive some simple, very specific predictions. While these circumstances are rarely encountered in detail, these situations provide approximate models which are useful in describing many real world circumstances.

What should we do when the forces which act are not so simple? What if they change with time, varying in both magnitude and direction? How can we hope to predict the motion of an object subject to complex forces like this? In fact these problems are not much more difficult than what we have done above, and our solution of them relies on the same fundamental approach. If we know what the acceleration of an object will be at every moment, we can predict its path as accurately as we like. Newton's second law tells us how to find the acceleration; we simply have to know the forces. This is how we determine the motion of objects, by combining Newton's second law with our knowledge of force laws, which allow us to predict the force which will act on an object given some knowledge of its properties and circumstances.

Combining these two ideas gives us 'dynamical equations'. The solutions of these equations represent all the motions which are possible. Consider the simple case of ballistic motion with no friction. After the object is launched, we need only combine Newton's second law with the appropriate force law:

$$\vec{F}_{\text{total}} = m\vec{a} \quad \text{and} \quad \vec{F} = -mg\hat{y}$$

$$m\vec{a} = -mg\hat{y}$$

This vector equation can be written as two separate equations, one describing the motion in each of the two directions, horizontal and vertical:

$$a_x\hat{x} + a_y\hat{y} = -g\hat{y}$$

$$a_x = \frac{d^2x}{dt^2} = 0$$

$$a_y = \frac{d^2y}{dt^2} = -g$$

These are two differential equations. The 'solutions' to these equations are any functions $x(t)$ and $y(t)$ which obey the relations above. Such solutions represent *all* possible motions of a real object under the influence of this force.

How might we find such a solution? When you take a course in differential equations, you will learn a variety of techniques which may help you to identify solutions, but in fact you can find them any way you like, including guessing.

In this case we need to find a function $x(t)$ which has a second derivative with respect to time equal to zero. Here is the most general form of such a function:

$$x(t) = A_x t + B_x$$

We can solve the dynamical equation describing the vertical motion in a similar way. For the vertical motion we need to write a function $y(t)$ whose second derivative with respect to time is not zero, but is instead $-g$. Here is the most general form for a function meeting that condition:

$$y(t) = -\frac{1}{2}gt^2 + A_y t + B_y$$

Notice that each of these equations has two currently unknown constants. Why are they present? As we have seen before, the equations of motion describe **all possible** motions. Since this is the case, they must have in their solutions flexibility sufficient to fit every particular case. Flexibility in the solutions appears as constants of integration in the solutions.

Given just the equations of motion, we can't say *anything* about the values of A_x , B_x , A_y , or B_y . These four constants can be determined *only* by reference to the particular details of our problem. What do these

constants mean? Remember that the function $x(t)$ represents the position of the object along the horizontal direction as a function of time. If we know the initial conditions for a particular motion, if we know the initial position and initial velocity in the horizontal direction, we can solve for these two constants.

$$x(t) = A_x t + B_x$$

$$x(0) = B_x \quad \text{so} \quad B_x = x_{\text{initial}}$$

$$\frac{dx(t)}{dt} = A_x \quad \text{so} \quad A_x = v_{\text{initial } x}$$

The same approach also works for the vertical motion.

$$y(t) = -\frac{1}{2} g t^2 + A_y t + B_y$$

$$y(0) = B_y \quad \text{so} \quad B_y = y_{\text{initial}}$$

$$\frac{dy(t)}{dt} = -g t + A_y \quad \text{so} \quad \left. \frac{dy(t)}{dt} \right|_{t=0} = v_{\text{initial } y} = A_y$$

So the constants which show up in our solutions are just the initial conditions of the problem. The dynamical equations which we get from Newton's laws and our knowledge of force laws tell us how objects can move, but they don't tell us how they *will* move. To know what will happen, we must first have some knowledge of the initial conditions, of how things start out. Newtonian mechanics has nothing to say about this, it can't tell you what the initial conditions are. Those come from some knowledge of the world. We can now rewrite our equations of motion in the familiar form:

$$x(t) = v_{\text{initial } x} t + x_{\text{initial}}$$

$$y(t) = -\frac{1}{2} g t^2 + v_{\text{initial } y} t + y_{\text{initial}}$$

We shouldn't focus too much on the notion that we need to know the *initial* conditions. In fact, one could find the constants not determined by the dynamical equations by knowing the position and velocity in the horizontal and vertical directions at any time. For instance, if we knew the position and velocity when the projectile landed, we could use the dynamical equations to learn where it came from. Specifying the conditions at any moment in the path would be enough to fully specify the motion of the object during the whole period in which the dynamical equations apply.

What to do with more complex force laws?

In some simple cases, the dynamical equations, which have a general form like:

$$\frac{d^2 \vec{s}(t)}{dt^2} = \frac{\vec{F}(\vec{s}, t)}{m}$$

will have solutions which are nice analytic functions like these. But what if they don't? What if the dynamical equations are too complicated to solve explicitly? When this happens (as it almost always does in realistic circumstances) we need to solve the dynamical equations using numerical methods.

Suppose we know the dynamical equations which govern the motion of an object. The solutions to these equations represent all possible motions which can happen given these circumstances. To find out which of these motions our object undergoes, we also need to know the 'initial conditions'. With these, we can figure out all the details of the motion, even if we can't analytically express it. Let's see how this works for a one dimensional motion, the use it to explore the more complex problem of projectile motion with friction. In one dimensional motion, the dynamical equation will take the form:

$$\frac{d^2x(t)}{dt^2} = a(t) = \frac{F(x,t)}{m}$$

If we know the force laws, we know $F(x,t)$. Now imagine we also know the initial position and velocity. Given these, we can also calculate the initial acceleration.

$$a(0) = \frac{F(x(0), 0)}{m}$$

We can use this information to predict the position and velocity of the object a short time Δt later. Here are equations that approximate this:

$$\begin{aligned} x(\Delta t) &\cong x(0) + v(0)\Delta t \\ v(\Delta t) &\cong v(0) + a(0)\Delta t \end{aligned}$$

Notice the approximate equalities here. These equations assume a constant velocity and constant acceleration throughout the time period Δt , something which is not precisely true. But so long as we keep the time period short compared to the times over which the motion of the object changes, these approximate relations can be as accurate as we like. Now that we know the new position and velocity, we can calculate a new acceleration and repeat this process, eventually predicting the whole motion of the object.

$$\begin{aligned} a(\Delta t) &\cong \frac{F(x(\Delta t), \Delta t)}{m} \\ x(2\Delta t) &\cong x(\Delta t) + v(\Delta t)\Delta t \\ v(2\Delta t) &\cong v(\Delta t) + a(\Delta t)\Delta t \end{aligned}$$

This is the essence of the numerical solution of dynamical equations. So long as the forces which act are known, this approach can be used to address incredibly complicated problems. The calculations involved are often very challenging, requiring an enormous number of steps, and hence very powerful computers. But this approach, in principle at least, allows us to predict motions any time the initial conditions and forces which act are well understood.

Numerical calculations of paths: real projectiles with friction

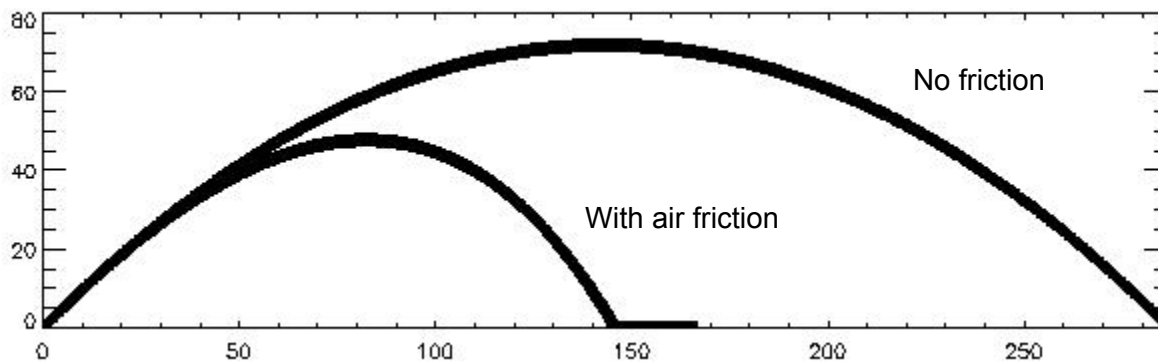
Real projectiles on Earth are never immune from air friction. They always feel a friction force acting opposite their direction of motion which increases as the velocity of the projectile increases. Exactly how the friction depends on velocity depends on the details. As outlined in Chapter 5, air friction will typically be proportional to the speed v and size if the object is small and moving slowly, or proportional to v^2 and area if the object is moving rapidly or is large. These two force laws are usually written:

$$F_{\text{slow small}} = 6\pi\eta r v \qquad F_{\text{large fast}} = \frac{1}{2} C \rho A v^2$$

In either case, friction acts in a direction opposite the motion. This force will slow the motion in *both* the horizontal and vertical directions, but the effect is qualitatively different for the two. In the horizontal direction, this resistance to motion can gradually reduce the horizontal velocity to zero, because there is no force in this direction available to keep the motion going.

In the vertical direction, friction will likewise act to resist motion. While an object is moving upward, friction acts in concert with the weight to slow the object more rapidly. As a result, the object will not rise as high as it would without air friction. Once the projectile turns around and starts to fall downward, the weight acts to increase the downward velocity while the air friction acts opposite this. Eventually, the object speeds up until the upward frictional force equals the weight, then it travels at “terminal velocity” we have discussed in the past.

What does the path of such a projectile look like? Instead of traveling horizontally at a constant rate, the horizontal motion slows, covering less and less distance in each unit of time. Meanwhile, the vertical velocity is everywhere slowed; the object doesn't rise as high and drops more slowly. We can calculate the resulting path using the numerical methods introduced above. They are sketched out for this problem in more detail below. The net result is a shorter, stubbier path, like this:



This example shown in this figure is a realistic representation of the path of a baseball launched from the bat of a power hitter. Without air friction, such a ball might travel close to 300 m, or around 1200 feet. With air friction, the ball only travels a bit less than 150 m, or close to 500 feet, something which would still qualify as a monster home run. Smashing it harder helps, but very much less than the range equation would suggest, because increasing the launch speed also increases the importance of the friction.

This path is calculated numerically. Let's look at the details a bit. In this case the dynamical equations are like those outlined for projectile motion above, but complicated by the presence of the friction force. Now we have:

$$\vec{a}(\vec{v}, t) = \frac{1}{m} \left(-mg\hat{y} - \frac{1}{2} C\rho A v^2 \hat{v} \right)$$

where with the symbol \hat{v} we mean a vector in the direction of the velocity at that point in time. We can now write equations predicting the motion a time Δt after the initial moment. They are a little more complex, because we have to determine the direction of motion at each step. We will express this as the angle $\theta(t)$ between the current velocity and the horizontal.

$$\sin(\theta(0)) = \frac{v_y(0)}{|\vec{v}|} \quad \text{and} \quad \cos(\theta(0)) = \frac{v_x(0)}{|\vec{v}|}$$

$$a_x(0) = -\frac{1}{2} C\rho A (v(0))^2 \cos(\theta(0))$$

$$x(\Delta t) = x(0) + v_x(0)\Delta t$$

$$v_x(\Delta t) = v_x(0) + a_x(0)\Delta t$$

$$a_y(0) = -g - \frac{1}{2} C\rho A (v(0))^2 \sin(\theta(0))$$

$$y(\Delta t) = y(0) + v_y(0)\Delta t$$

$$v_y(\Delta t) = v_y(0) + a_y(0)\Delta t$$

This process is then repeated to find the position and velocity in both horizontal and vertical motion at each step, and the path is traced out. So although we cannot write an analytic equation which expresses the path of a projectile which experiences air friction, we can use the dynamical equations to predict such motion, and study its dependence on conditions, with arbitrary accuracy.

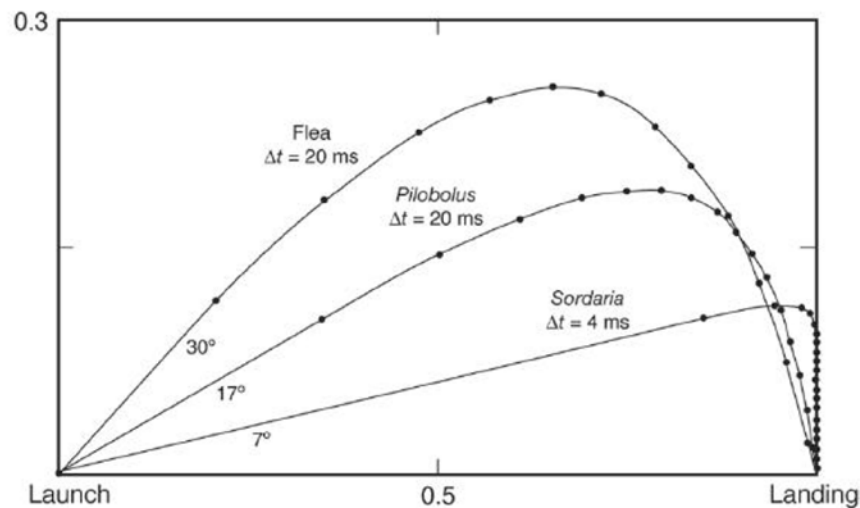
There are a surprising number of natural projectiles, things which acquire all their velocity at launch and then fly ballistically through the air. Some are whole organisms, from gazelles and kangaroos to gibbons and high jumpers. Many are the seeds of plants and spores of fungi, fired into the air by a variety of mechanisms intended to spread them beyond where they will necessarily compete with their parent. While most of the whole organisms and some of the seeds fly in the large-fast friction regime, many of the seeds and all of the spores are very small indeed, and definitely experience slow-small friction. In almost every case, friction plays a significant role, often severely limiting the maximum possible range.

When friction is present the launch angle which gives the maximum range for a given launch velocity is always less than 45° . You can see why if you recognize that horizontal velocity is always being lost when air friction is present. Because of this, the horizontal velocity late in the flight is less than the horizontal velocity early in the flight. Aiming a little lower allows the projectile to cover a lot of horizontal ground at the start, while the horizontal velocity remains large.

One way to quantify the impact of friction on a projectile's motion is to consider the range penalty. If you compare the maximum range with friction to the maximum range without friction, you get an interesting measure of the importance of friction:

$$\text{range penalty} = \frac{r_{\text{no friction}} - r_{\text{friction}}}{r_{\text{no friction}}}$$

An excellent summary of natural projectiles is provided in a *Journal of Biosciences*ⁱⁱ article by Steven Vogel. In this article, Vogel shows that this range penalty varies from as little as 1.1% for a jumping Kangaroo Rat to 99.997% for the very tiny spores fired from the fungus *Giberella zeae*. As friction becomes more important, the optimum launch angles fall lower and lower, as shown in this figure taken from Vogel's paper.



This figure shows the optimum launch angles for three small natural projectiles. The first is a flea, a jumper about 1.6 mm in size who launches herself at about 3 m/s. This is a pretty amazing speed; unimpeded, she would travel nearly 2000 times her own size in the first second. The second example is the sporangium of a typical member of the *Pilobolus* genus of fungi, which grow on herbivore dung. To propagate, they must get their spores out of the dung and on to some fresh vegetation, where they can be eaten by a new herbivore. They do this by building up quite substantial fluid pressure behind the sporangium, then letting it burst forth, blasting off toward its target. The third example shows the very tiny groups of spores launched by the *Sordaria* genus of fungi, which also live in herbivore dung. Despite the very tiny size of these spore groups, only about 40 μm in diameter, they are launched at speed of up to 30 m/s, and despite this very high launch velocity, friction resists their travel to less than 10 cm.

The flea pays a drag tax of about 38%, the *Pilobolus* about 98%, and the *Sordaria* almost 99.96%. You can see from the figure above the increasing importance of friction in these three paths. All are clearly asymmetrical, falling more steeply than they rise. On the figure above, each path is marked off at regular intervals of time, to give you a sense of how dramatically the velocity of each is reduced, beginning immediately upon launch. Many organisms which fire the smallest projectiles effectively give up on

distance, often launching their tiny missiles straight up instead, hoping to loft them up where a passing breeze might be able to give them a ride.

Numerical calculation of paths: planetary orbits

Another problem which begins simply, then quickly becomes more complex is the orbit of planets. The circular orbit of a planet around a single star is simple, and we have analyzed it in an earlier chapter. In this case, the centripetal force is provided by the gravitational pull of the star. Newton's universal law of gravity combines with what we know about circular motion to guarantee a simple relation between distance from the star and orbital period. If the orbit is not circular, the problem is more complex, but so long as the only force on the planet comes from the star, it is still analytically soluble. In fact, Newton solved it himself, showing that the planets usually travel in elliptical orbits, with the Sun resting at one focus of each planet's elliptical orbit.

But the real solar system is more complex still. While the largest force acting on each planet comes from the Sun, each is also tugged by all of the other planets and moons in the solar system. These extra forces 'perturb' the orbits of the planets in a variety of small, but perfectly measurable ways. The more complex motion which results can be predicted, but doing this requires tracking the positions and motions of all the planets. In each time step, the total force on each planet is determined from its gravitational interaction with both the Sun and all of the other planets. Given this acceleration, new positions and velocities for the next time step are calculated, and the process begins again. Though the number of objects in the solar system is not large, the timescales which ought to be considered are. This makes calculating the future of the solar system a serious numerical challenge. The best current predictions suggest that the solar system will likely remain stable for the next few billion years, but in truth the predictions become increasingly uncertain after the next 100 million years or so.

Numerical calculations of paths: N-body calculations

Many real problems in the world involve complicated interactions of a large number of particles. These problems are generally called 'N-body' calculations, because they involve some number (N) of individual parts. The number of parts N might vary from three to tens of billions. Good examples of N-body problems include questions like the fate of the solar system ($N \sim 10$), the folding of a protein like hemoglobin ($N \sim 2 \times 10^4$), or the flow of a fluid ($N \sim 10^6$).

To study these N-body problems you need to track the position and velocity of each particle as a function of time:

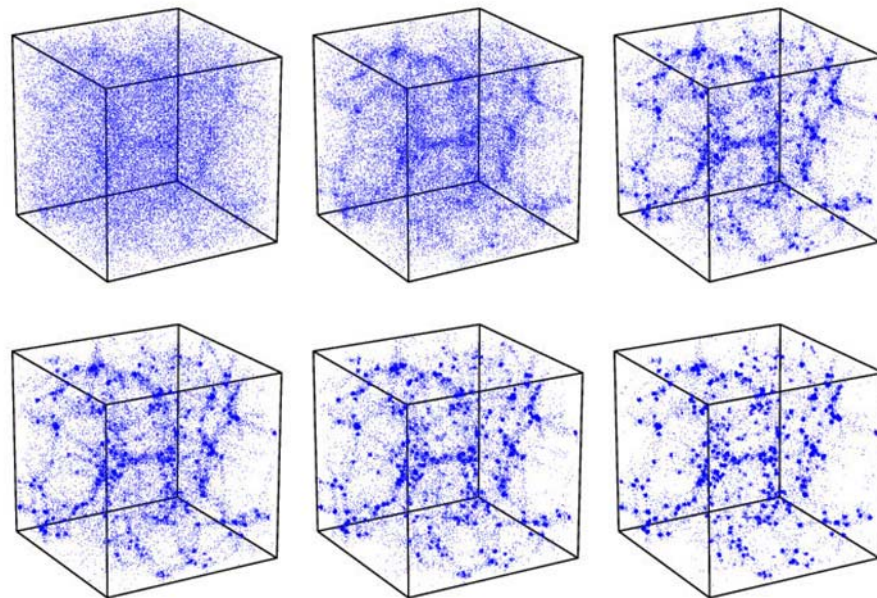
$$x_i(t), y_i(t), z_i(t), v_x^i(t), v_y^i(t), v_z^i(t)$$

Given these quantities, you can use the appropriate dynamical equations, finding the force on each particle by summing up the forces exerted on it by all the other particles:

$$\vec{F}_i(t) = \sum_j \vec{F}_{ij} = m\vec{a}_i$$

Of course you'd also have to know the initial conditions for each particle as well. For each step, you'd have to loop over all the particles and calculate the new positions, velocities, and accelerations (from the forces). Once you can do this, you just continue for as many steps as you need.

Perhaps the most dramatic form of N-body simulation aims to track the growth of structure throughout the entire universe, the so-called 'cosmological n-body' simulations. We now know that the universe began in a nearly uniform state, with a matter density almost – but not quite – the same everywhere. Modern cosmology simulations begin by spreading perhaps ten billion particles nearly uniformly in space, giving each small, random velocities. After this, they track the gravitational forces on each particle and follow them, step by step, along the paths they follow. They do this for the full 13.7 billion year age of the universe, and sometimes even into the distant future. Such simulations show how an initial overdensity grows with time, gravitationally pulling in other particles around it, until eventually galaxies and clusters of galaxies form.



The figure above shows a series of snapshots from such a cosmological simulation. The box on the upper left shows the early universe, in which the matter density remains nearly constant. Each subsequent box shows the universe at later and later times, and you can see how gravity, pulling things together, gives rise to the formation of cosmic structure.

A Quick Summary of Some Important Relations

(Not including material from section 8.3, which is relevant only to the Physics 136 Lab)

First model of simple 2D motion – uniform circular motion:

An object traveling in a circle at a constant speed experiences an acceleration with a fixed magnitude, always pointing toward the center of the circle. Forces must be applied to create this acceleration.

$$a_{\text{centripetal}} = \frac{v^2}{r} \quad \Sigma F = \frac{mv^2}{r}$$

Remember that there is no new ‘centripetal force’. When you see uniform circular motion, you just know that the total force on the object must have this magnitude and point toward the center of the circle.

Second model of simple 2D motion – a constant force (projectile motion):

An object experiencing one force constant in magnitude and direction travels in a parabola opening in the direction of the force. When air friction is ignored, projectiles can be modeled by this motion. In this case, the vertical and horizontal motions are independent. The horizontal motion is uniform, while the vertical motion is accelerated by gravity.

$$v_x = v_{0x}$$

$$x_f = x_0 + v_{0x}\Delta t$$

$$v_y = v_{0y} - g\Delta t$$

$$y_f = y_0 + v_{0y}\Delta t - \frac{1}{2}g(\Delta t)^2$$

To apply these, you will often use one motion to find, for example, how long the object is in the air, then use this information in the other to find out, for example, how far it travels. For the simple case of a projectile launched from the ground which returns to the ground, this approach yields a range equation:

$$\text{Range} = \frac{v_0^2 \sin(2\theta)}{g}$$

More complex motions:

When neither of these conditions is met, the 2D and 3D motion of an object can still be predicted, so long as the forces which act on the object are known at all times. This prediction is done by considering the motion step-by-step.

ⁱ Van Der Burgt, X., 1997, *Journal of Tropical Ecology*, **13**, 145.

ⁱⁱ Vogel, S., 2005, *Journal of Biosciences*, **30**, 167.

9. What's happening: work and kinetic energy, the scalar quantity of motion

- 1) Interactions, systems, and state
 - i. The energy concept: a little history
- 2) Work and the energy of motion
 - i. Work and changes in kinetic energy contrasted to impulse and changes in momentum
- 3) Calculating work and determining changes in energy
 - i. Some simple examples
 - ii. The word "work" in physics and the vernacular
 - iii. Energy is relative, but changes in energy are absolute
- 4) Using the work-energy theorem
 - i. Constant forces applied to objects traveling in straight lines
 - ii. Comparing the impulse-momentum and work-energy theorems
 - iii. Changing forces applied to objects traveling in straight lines
 - iv. Work and objects traveling in curved paths
 - v. How rapidly work is being done: power
- 5) Energy and rotational motion
 - i. Torque, angular displacement, and work
- 6) Quantifying energy in technology and life: some interesting comparisons

Physics for the Life Sciences: Chapter 9

9.1 The energy concept: a little history and context

The physics we have discussed so far is the physics of Newton, expressed in his famous laws. They are largely concerned with force and motion. From the time of the publication of Newton's work in 1687 until the beginning of the 19th century, force and motion played the central role in physics. This was a very successful approach; it explained the motions of planets and of objects thrown through the air, understood the influence of friction and weight, and noted the way in which motion in curved paths *requires* the presence of forces. But of course it left most of the world unexplained.

Newton's physics has nothing to say about the thermal and chemical changes which take place when you cook. It can't explain the wind or rain, why the Sun shines, or wood burns. Most important, it has nothing at all to say about life and how living things make their way in the world. For Newton, life remained as magic as ever, utterly different from inanimate matter. The fates of mighty planets were governed by physical laws, but something as mundane as the growth of grass or the leap of a kitten was not.

Beginning in the early 19th century, a number of scientists began to expand the content of physics to include a broader range of phenomena. Many factors encouraged this, but most especially the explosion of industry which accompanied the invention of steam engines. These devices converted heat, generated by burning wood or coal, into mechanical motion. They provided motive force, previously available almost exclusively from living things. Steam engines could move things around, just as horses, oxen, and people could; converting their fuel into motion. Such engines could even move themselves, bringing 'to life' steam trains, ships, and cars. Active motion was no longer the sole province of life. Focused study of

these engines, motivated to a great degree by purely practical concerns, played the central role in developing the science of energy.

Eventually, it became clear that Newton had missed a major element of physics. He had emphasized the importance of momentum, and gave us the tools to describe its change. But he had not recognized the parallel importance of energy. Today we view momentum and energy as coequal concepts, each expressing essential aspects of a system. Momentum changes when an impulse is applied, through application of a force for a period of time. Energy can change in more than one way. It can be altered by work; the application of a force through some displacement. But it can also be altered by a transfer of heat.

In this chapter we will focus on a first form of energy; the kinetic energy found in the motion of an object. We will see in detail how this form of energy is changed by work. In the next chapter, we will explore a second form of energy; the potential energy associated with interactions. These first forms of energy are related to Newton's mechanics in an obvious way; so obvious indeed that they seem just another way of seeing what Newton already revealed. The real power of the energy concept lies in its ability to include not only the obvious energy revealed by the motions and configurations of objects, but the much less apparent thermal energy contained within them.

Once we have explored the macroscopic forms of energy, we will spend several chapters understanding how objects possess internal thermal energy. In Chapter 11 will explore collisions, and see how they allow atoms in a gas to share energy. In Chapter 12 will see how the atoms in solids store energy both in motion and by stretching the bonds which hold them in place. Finally in Chapter 13 we will fully explore how the invisible motions of atoms in a material are revealed by its temperature. This connection will show us how the thermal energy harnessed by steam engines, and indeed by life, is related to the motions of atoms.

9.2 Work and the energy of motion

From the beginning, there were thoughtful people who realized there must be more to change in the world than just force and momentum. Christian Huygens, a brilliant Dutch contemporary of Newton, was very aware of this. He recognized, for example, that when a bomb goes off and pieces fly in every direction, no net force has acted. All the forces are internal; when one part of the bomb pushes on another, the second pushes back with an equal and opposite force on the first. So somehow, though there was no net force applied to this system, there was clearly a *very* big change in motion.

Huygens saw there must be another measure of motion not adequately expressed by forces and momentum. There are many ways you can have no net force on a system, and hence no change in total momentum, and yet have substantial changes in motion. To quantify these changes, a new *scalar* measure of motion was required. Huygens wanted to invent a 'quantity' of motion which describes in a general way whether things are moving or not, without regard to the direction they're traveling. He adopted for this quantity the Latin name "*vis viva*", or the motion of life. Today we refer to this scalar measure of motion as "*kinetic energy*". In the sections that follow, we will see how to quantify kinetic energy and the work which is responsible for changes in it.

As usual in science, the notion that another important concept was lurking about was apparent not only to Huygens. Others had glimpsed it before, including the Greeks, and (of course) Galileo, particularly in connection with the simple machines like levers and inclined planes. Simple machines allow you to do something which would usually require a large force acting over a short distance by instead applying a small force over a long distance. That the same goal could be accomplished either way suggested that the product of force times distance was important.

The product of force times distance was finally codified as important by Thomas Young, a medical doctor of remarkable mental agility in the early 19th century. He called this quantity (force times distance) "work", and suggested that work is what changes the vis viva, or energy of motion in an object. The work defined by Young quantifies the energy transferred from one object to another through the action of forces.

In Young's initial example, the energy transferred to (or from) the object appears in it as the kinetic energy of motion, and this is the first example we will develop. But the energy transfer quantified by work can appear in an object in many other forms, for example as the energy associated with tension, compression, or temperature. In every case though, the work calculated according to Young's definition correctly quantifies the energy transfer associated with the action of a force.

It is interesting that someone outside the usual path of physicists made this discovery. The problem with understanding work and energy had little to do with mathematical complexity. The calculations required are simple, but the ideas are subtle and abstract. They are also very much questions of everyday life. Perhaps this is why so much of the progress in the study of energy came from people with varied interests; from doctors like Young and Julius von Mayer, engineers like William Rankine and Sadi Carnot, and even brewers like James Joule.

Interactions, Systems, and State

By introducing energy we will undertake a substantial expansion of our topic, putting together tools which allow us to discuss a much broader range of phenomena. As we embark on this, it is useful to think generally about what we're doing.

In the most general sense, what we want to describe in physics is how one thing affects another. We have so far concentrated on a few simple effects one object can have on another: one object might make another accelerate, or it might make it deform. We invented a name for the means by which one object creates acceleration or deformation in another; we called this a 'force'. But there are many other ways in which objects affect one another. If I place a hot object next to a cool one, the hot one cools off while the cool one heats up. If I place a metal in acid, it dissolves, and hydrogen gas is released. In the most general way, we can speak of all these 'effects' as interactions.

The interactions we have been concerned with so far, those associated with forces, we might call mechanical interactions. Those involving heat we call 'thermal'. Interactions which give rise to changes in chemical composition we call 'chemical'. When you rub a balloon on the rug and then stick it to the wall we speak of 'electrical' interactions. Magnets stick to refrigerators through 'magnetic' interactions. Often what we see, especially in living things, is a mix of many of these. If I plant a seed in the ground and water it, a tree can grow. This complicated effect is built of many different kinds of interactions;

chemical interactions within the cells of the plant, electromagnetic interactions with light from the Sun, and mechanical interactions as the tree pushes through the soil, raises itself against the pull of gravity, and resists the strain of windstorms. The rich complexity of life emerges from the combination of many essentially simple interactions.

In each case we know an interaction has occurred because we see a change in the properties of the systems involved. The cold object became hot, the metal dissolved. There is always clear, repeatable evidence that something has happened. The interactions which concern us in physics are brought to our attention through the clear and repeatable effects which they have on the world around them.

This is an important point, because people have always been fascinated by apparent interactions which *don't* bring about clear and repeatable effects; things like astrology, in which the positions of the planets are thought to influence the course of one's life. These subjects are outside the bounds of science not because of prejudice against them, but because they don't create the kind of repeatable, reliable effects on their surroundings which make it useful for science to discuss them. If they did, scientists would surely be eager to study them, because discovering something truly new is the most exciting and highly rewarded achievement in science.

To speak generally about interactions we need to refine what we mean when we talk about the 'things' which are interacting. To use a general term, we'll call these 'systems'. In many situations we have analyzed so far, the 'system' is one of the objects we would have drawn a free body diagram for; a block or a ball, a monkey or an elephant. Note that the definition of a system is quite arbitrary. We get to draw the lines around these systems where we want. So while a system might be just one atom, it might also be the whole gas which fills a room, or the whole Solar system. It is only important that we are careful to be clear about what we are discussing in any particular case.

The way in which we perceive interactions between systems is through changes in their properties. Typical mechanical properties include momentum, shape, and mass. Systems may have other properties we have so far not discussed, like temperature, pressure, chemical composition, electric charge, magnetization, and physical state (solid, liquid, gas, or plasma). For a given situation, we can describe the system by listing the values for all the relevant properties which describe it. In a certain sense, the system *is* the set of properties which describe it. If there were something more needed to describe it, we would simply add that to the description of properties.

So, to do physics, we need to talk about **interactions** between **systems**, which alter their **state**. We've already done a lot of this for mechanical interactions between objects which alter their momenta or shapes.

To achieve this more general description of the world, it's clear that we won't be able to describe everything in terms of forces and accelerations. We must begin to include more. In this chapter we take a first step down that path, as we begin to introduce the idea of energy. We will find that 'energy' is another property of a system, something which we must know to describe its state. It is a very important, highly general thing, and especially essential in understanding life. So we will spend the next several chapters focusing on the idea of energy, and come to see life as a particularly beautiful element of the natural flow of energy from the Sun down onto the Earth, and eventually back off into space.

Work and changes in kinetic energy contrasted to impulse and changes in momentum

We will begin by considering the motion of a point object, and determining what work done on such an object implies. We have previously recognized that when a force acts for some period of *time*, it generates an 'Impulse'. Impulse changes the momentum of an object:

$$\int \vec{F} dt = \Delta \vec{p}$$

To calculate impulse, we sum the product of a vector force with a scalar time. This gives a vector result, which is the vector change in the momentum. This is the aspect of changing motion that Newton clearly understood. In fact the impulse-momentum theorem is just a version of Newton's second law.

Young recognized that there was also something important about the product of a force \vec{F} and the distance $d\vec{s}$ through which an object travels while the force acts. There is a complication though. Both the force \vec{F} and the displacement $d\vec{s}$ are vectors, which may or may not be in the same direction. What matters for work is the extent to which the force acts along the direction of motion: a quantity represented by the scalar product of force and displacement:

$$\int \vec{F} \cdot d\vec{s} = \text{'Work'}$$

The total work done on an object is found by adding up the component of the force along the direction of motion times the displacement as the object moves. Only forces along the direction of motion do work, and work is done only when the object moves! The units of work are Newton*meters, or kg*m²/s². This unit (1 kg*m²/s²) is called a "Joule". Since work is a measure of the transfer of energy, these Joules are also the units of energy.

Given this definition for work, let's find out what happens to a point object when work is done to it. To keep things simple, begin with a case where the force acts completely along the direction of motion.

$$\int_i^f \vec{F} \cdot d\vec{s} = \int_i^f F ds = \int_i^f m a ds$$

We can rewrite the acceleration in this equation using the chain rule of calculus:

$$a = \frac{dv}{dt} = \frac{dv}{ds} \frac{ds}{dt} = v \frac{dv}{ds}$$

Inserting this in the equation for work above, we find:

$$\begin{aligned} \int_i^f m a ds &= \int_i^f m \left(v \frac{dv}{ds} \right) ds = \int_i^f m v dv = \frac{1}{2} m v^2 \Big|_i^f \\ &= \frac{1}{2} m v_f^2 - \frac{1}{2} m v_i^2 = \Delta \left(\frac{1}{2} m v^2 \right) = \Delta KE \end{aligned}$$

Adding up the scalar product of \vec{F} and $d\vec{s}$ gives a result which is the change in a scalar quantity. If the force acts on a point object, this scalar quantity describes the overall amount of motion, ignoring its direction, and is called kinetic energy:

$$KE = \frac{1}{2}mv^2$$

Just as the product of the force and the time over which it acts is the change in momentum, the product of the force and the *distance* over which it acts is the change in kinetic energy.

Impulse-momentum theorem $\int \vec{F} dt = \Delta\vec{p}$ changes the vector quantity of motion

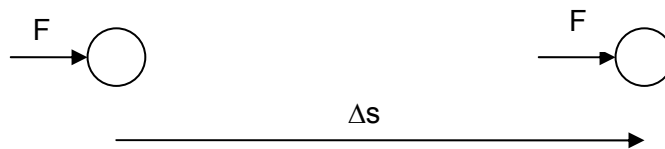
Work-energy theorem $\int \vec{F} \cdot d\vec{s} = \Delta KE$ changes the scalar quantity of motion

What is kinetic energy? Kinetic energy is one of the many forms of energy which a system might possess. It is the energy of motion, a scalar measure of whether an object possesses motion at all, independent of what direction it might be in. When we ask how much energy a system has, we quantify its energy of motion by calculating its kinetic energy.

We have found here that, for a point object at least, changes in kinetic energy are determined by calculating the work done on the object. This is completely parallel to our discovery that impulse and changes in momentum are related. If you want to know how the momentum of a system changes, calculate the impulse applied to it. If you want to know how the kinetic energy of a system changes, calculate the work done on it. Work represents a transfer of energy, just as impulse represents a transfer of momentum.

9.3 Calculating work and determining changes in energy

Consider first the case of a force applied to an object along its direction of motion. Nothing fancy here, just a constant force accelerating an object from rest.



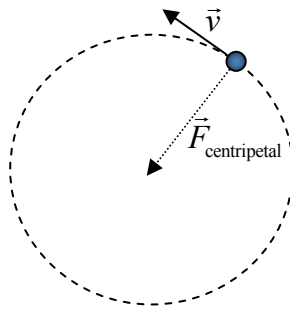
In this case the work done by the force is just the product of the force times the displacement:

$$\int \vec{F} \cdot d\vec{s} = F \Delta s$$

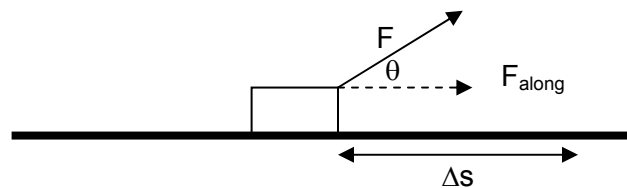
Clearly such a force acting in this way affects a change in the motion. Since the force acts wholly along the direction of motion, it does positive work; it increases the kinetic energy.

What if the force is opposite the direction of motion? In this case, the scalar product of force and displacement is negative, and the work done on the object is negative. Negative work implies a reduction of kinetic energy. To repeat, work is positive if the force is in the direction of motion, and negative when the force is opposite the direction of motion.

What happens if there is a force which does not act purely along the direction of motion? Imagine an object traveling around in a circle. Are there any unbalanced forces acting on it? There must be; it's traveling in a circle. Do these forces change the motion of the object? Yes, they alter the direction of its motion. Do they do any work? No, because the direction of motion is always perpendicular to the direction of the force. The object never moves along the direction of the force. This is a beautiful example, because the force is not zero, and the motion does change, but the work ($F_{\text{along}}\Delta s$) is zero, and so in this case the change in energy is also zero.



What about an intermediate case, in which the force is *partly* along the direction of motion?



In this case $F_{\text{along}} = F \cos(\theta)$ where θ is the angle between the force and the direction of motion. The work done here (and hence the change in kinetic energy achieved) is intermediate between the two cases above; neither as large as the maximum possible ($F\Delta s$) nor as small as the minimum the force could do while the object moves (zero). In this case the work is:

$$W = F_{\text{along}}\Delta s = F \cos(\theta)\Delta s$$

The word "work" in physics and in the vernacular

The concept of work within physics is often confusing when first encountered. Much of the confusion stems from familiarity. We use the word 'work' all the time in everyday life, where it has a familiar, though vague meaning. But now we want to take this broad and general meaning, the meaning we use in

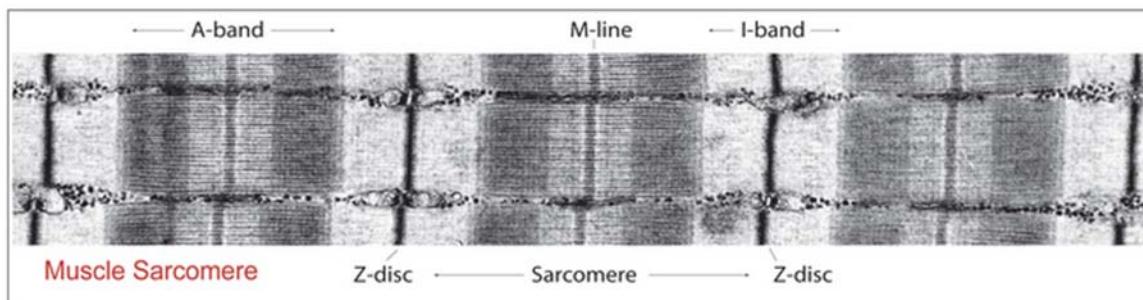
everyday life, and replace it with something **very** specific and exact. You must remember this when you think about work in physics.

Consider a simple task: carrying a suitcase through the airport. The force you apply is always perpendicular to the direction of motion; the force is upward, while the motion is horizontal. As a result, you do no work on the suitcase at all. Nevertheless, you'll probably get very tired carrying it from one terminal to another.

Work in physics and the work you usually talk about are clearly not the same. Work is done on an object in physics only when a force acts along the direction of motion. You're probably more inclined to think of work as something which wears you out. How can physicists say that something which exhausts you doesn't involve work? Here's a clue that the physicists just might be right. If you put that same suitcase on a cart with nice, low-friction wheels, you can move it through the same airport with almost no effort. The suitcase moves in the same way; what happened to it was the same. But in the first case you were sweaty and exhausted, in the second you were not. So perhaps the physicists are right about no work being done on the suitcase, and we should examine instead what happened to you.

The solution to this mystery lies in the biology of muscle. The muscles of vertebrates are made of many tiny parts assembled in a hierarchical way. The sarcomere is perhaps the fundamental unit, a little $\sim 2 \mu\text{m}$ long assembly of myosin and actin (two proteins) capable of contraction. When it contracts, it applies a force to anything attached to it. A myofibril is made of perhaps 10,000 sarcomeres stacked end to end is a few centimeters long. The myofibril contracts when all of its sarcomeres do. A group of a thousand or so myofibrils connected in parallel make up each muscle fiber. And a muscle like your calf muscle is made of a million or so muscle fibers lined up side by side.

The sarcomeres can supply a force only by contracting. In the skeletal muscle you use to carry a suitcase (for example), each sarcomere undergoes a short, strong contraction, then takes some time to relax back to its original length; it twitches. To create a sustained force, groups of 500 or so muscle fibers twitch over and over again. The times when they twitch are staggered, producing an approximately constant combined force. To hold your suitcase at rest, the fibers in your muscles must contract repeatedly, 40 or 50 times a second. Each time they contract, they apply a force *along the direction of their motion*. So each twitch really does involve doing work! They do not, however, do any work at all on the suitcase. Instead, the work your muscle does while supporting a static load is done on you. The energy associated with this work never makes it into kinetic form, but instead is converted quickly into heat. All that work, the exhausting task of carrying the suitcase, really does nothing more than convert some of your chemical energy into heat.



Energy is relative, but *changes* in energy are absolute

There is a small but essential point lurking here. What we care about in physics is how energy *changes*, not what its absolute value is. You can guess that this might be true from the definition of kinetic energy: $KE = \frac{1}{2}mv^2$. We have seen how the velocity of an object can be different when viewed by different observers moving relative to one another. Two such observers comparing the total energy of any object would clearly disagree. As an example, imagine a passenger in a car (speeding along at 70 mph) who reaches out the window and drops a beer bottle. To the passenger, the energy and momentum of the bottle is zero. To the pedestrian watching from the sidewalk, the energy and momentum of the bottle is dangerously large. Who is right?

In fact they both are! Nothing in physics depends on the absolute value of the energy or momentum. To see why, consider what would happen if the bottle, just after being released, struck a rock along the road. The moving observer would see the bottle hit the rock and zoom off behind her, undergoing a sudden, large change in momentum. The stationary observer would see the bottle zoom toward him, shatter on the rock and stop moving forward, undergoing a sudden, large change in momentum. They would both see the same change in momentum and the same change in energy, and hence infer the same force between the bottle and the rock. This is why the impulse-momentum and work-energy theorems are only concerned with changes in momentum and energy, and not their absolute values.

9.4 Using the Work-Energy Theorem: simple examples

We can use the work energy theorem in a variety of ways. For example: If we know the force which acts on a system and the distance through which it travels, we can calculate changes in energy:

If I shove a 3 kg book across the table and it slides 2 m in coming to rest, how might we estimate its initial speed? To solve this, we will use the work energy theorem. If we calculate the work done by friction, we can find the change in kinetic energy, and from this infer the initial velocity.

The force which slows the book is kinetic friction. How large will it be? To calculate this, we would need to know the coefficient of kinetic friction between the book and the table. Remembering back to Chapter 5, typical values for friction between ordinary dry surfaces like these are around $\mu_k \approx 0.3$, so let's assume that. For this book, this implies a frictional force:

$$F_{\text{kinetic}} = \mu_k F_{\text{normal}} = 0.3 \times 3 \text{ kg} \times 9.8 \frac{\text{m}}{\text{s}} \approx 9 \text{ N}$$

We know that the kinetic friction force will be directly opposite the direction of motion at all times, so that the total work done by friction can be easily calculated:

$$W = \int \vec{F}_{\text{kinetic}} \cdot d\vec{s} = -F \Delta s = -9 \text{ N} \times 2 \text{ m} = -18 \text{ J}$$

Applying the work-energy theorem gives us:

$$\begin{aligned}
 W &= \Delta KE = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 \\
 -18 \text{ J} &= -\frac{1}{2}mv_i^2 \\
 v_i &= \sqrt{\frac{2 \times 18 \text{ J}}{3 \text{ kg}}} = 3.5 \frac{\text{m}}{\text{s}}
 \end{aligned}$$

Here is another example using the work-energy theorem to find the size of a frictional force.

If we know how much the energy of a system changes and we know how far it travels, we can estimate the net force which acts on it. Imagine that you're asked to estimate the size of the coefficient of kinetic friction between a child's sled and the snow. If you had a way to measure the sled's velocity, you might measure this when it reaches the bottom of the hill, see how far it slides on level ground before coming to a halt, and use this to estimate the size of the frictional force it experiences.

$$\begin{aligned}
 W &= \int \vec{F}_{\text{kinetic}} \cdot d\vec{s} = -F_{\text{kinetic}} \Delta s \\
 F_{\text{kinetic}} \Delta s &= \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 = \frac{1}{2}mv_i^2 \\
 \mu_k mg \Delta s &= \frac{1}{2}mv_i^2 \\
 \mu_k &= \frac{v_i^2}{2g\Delta s}
 \end{aligned}$$

Let's say the sled arrives at the level base of the hill traveling at 5 m/s and comes to a halt while sliding through a distance of 20 m, what is the coefficient of kinetic friction between the sled and the snow?

$$\mu_k = \frac{(5 \frac{\text{m}}{\text{s}})^2}{2 \times 9.8 \frac{\text{m}}{\text{s}^2} \times 20 \text{ m}} = 0.06$$

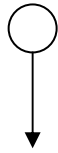
These examples should give some idea of how you can use the connections encoded in the work-energy theorem to relate forces to observed changes in motion in new ways.

Comparing the work-energy and impulse-momentum theorems

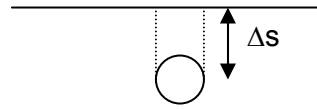
Both the work-energy and impulse-momentum theorems connect forces to changes in motion. It's useful to compare the two in some simple circumstances.

A ball is dropped from some height into a pile of sand. We know, from our earlier work, how to estimate the speed with which the ball arrives at the ground. Let's call this speed v_i . Once the ball hits the sand, it comes to a halt after penetrating some distance Δs :

Before



After



How does this interaction between the ball and the sand look from the perspective of the impulse-momentum and work-energy theorems? First let's look at the impulse-momentum version. Since we know the change in momentum, we can find the total impulse:

$$\text{Impulse} = \int \vec{F} dt = \Delta \vec{p} = \vec{p}_f - \vec{p}_i = -m\vec{v}_i = mv_i \hat{y}$$

From this we can find the average force which acts on it while it is coming to a stop. We get this by estimating the time of interaction Δt_{stop} :

$$\begin{aligned} \text{Impulse} &= \vec{F}_{\text{average}} \Delta t_{\text{stop}} = mv_i \hat{y} \\ \vec{F}_{\text{average}} &= \frac{mv_i \hat{y}}{\Delta t_{\text{stop}}} \end{aligned}$$

What is Δt_{stop} ? We might estimate it by saying $\Delta t_{\text{stop}} = \Delta s / v_{\text{average}}$, where v_{average} is the average velocity of the ball while it is slowing down. We could estimate this as $v_{\text{average}} = v_i / 2$, and obtain:

$$\Delta t_{\text{stop}} = \frac{\Delta s}{v_{\text{average}}} = \frac{2\Delta s}{v_i}$$

Putting these together, we get a final estimate of the average force, obtained from the impulse-momentum theorem, of:

$$\vec{F}_{\text{average}} = \frac{mv_i^2}{2\Delta s} \hat{y}$$

Now let's consider exactly the same situation from the perspective of the work-energy theorem. In this case, we know the change in energy, and want to find the force which does the work:

$$\text{Work} = \int \vec{F} \cdot d\vec{s} = \Delta KE = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 = -\frac{1}{2}mv_i^2$$

Here the force acts opposite the direction of motion, so that:

$$\int \vec{F} \cdot d\vec{s} = -F \Delta s$$

and from this we can find an average force:

$$F_{\text{average}} = \frac{mv_i^2}{2\Delta s}$$

Notice that these two estimates, one from the impulse-momentum theorem and one from the work-energy theorem, are completely consistent, as indeed they must be. Now let's consider a slightly more complex example: an elastic collision between a ball and the floor.

A ball falls downward, arrives at the floor with speed v_i , and bounces back upward. These are the initial and final states:



In this case the Impulse is:

$$\text{Impulse} = \int \vec{F} dt = \Delta \vec{p} = \vec{p}_f - \vec{p}_i = mv_i \hat{y} - (-mv_i \hat{y}) = 2mv_i \hat{y}$$

If we could estimate the time during which the ball is actually in contact with the floor as Δt_{bounce} we might write:

$$\vec{F}_{\text{average}} = \frac{2mv_i}{\Delta t_{\text{bounce}}} \hat{y}$$

What happens if we redo this using the work energy approach:

$$\text{Work} = \int \vec{F} \cdot d\vec{s} = \Delta KE = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 = 0$$

What's going on here? To understand we have to think about how the ball actually moves *during* the collision (which lasts for a total time Δt_{bounce}):



The ball reaches the floor (shown on the left), squashes down until it comes to rest (in the center), then springs back in the opposite direction (on the right). The *force* which alters the direction of the ball is always acting upward. The motion Δs is first downward (as the ball squashes) and then back upward (as it springs back). As a result this force first does negative work on the object, reducing its KE to zero, and then does positive work on it, restoring its KE.

So in this more complex case as well, both viewpoints provide a consistent picture of the interaction. This will of course always be true. Learning to see how this is true in various situations will refine your understanding of the work-energy and impulse-momentum theorems, and should help you to apply them in a variety of circumstances.

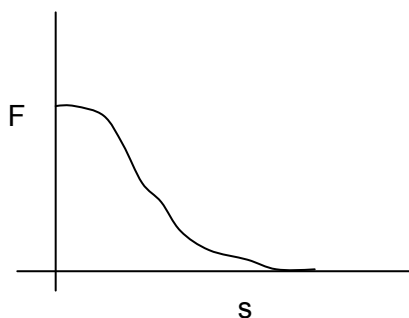
Changing forces applied to objects moving in straight lines

Calculating the work done by a constant force on an object moving in a straight line is simple. Now we want to consider two variations. First, what is the work done by a varying force on an object moving in a straight line?

$$\text{Work} = \int \vec{F}(s) \cdot d\vec{s} = \int F_{\text{along}}(s) ds$$

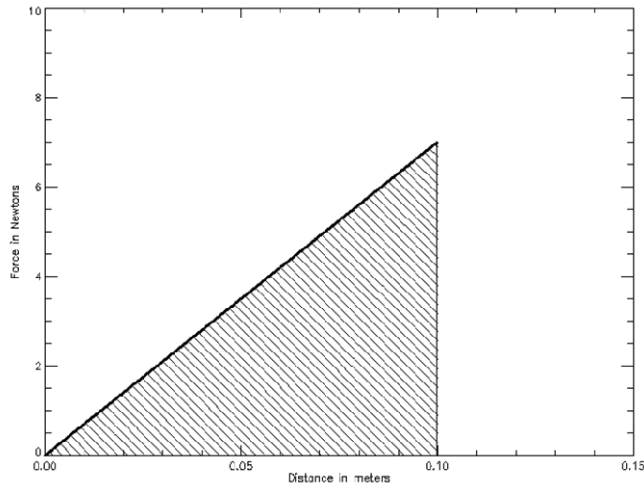
The work done as the object travels through each little distance is just the force at that position multiplied by the distance traveled. So if we imagine a graph depicting the force applied at each position during the motion, we can calculate the work done by finding the area under that graph.

Consider a simple example. Imagine that you are helping a child to learn to ride a bike. If you give them a push to get them started, you will start out pushing hard, and gradually decrease your force as the child accelerates away. So the force vs. distance curve for this might look like:



The total work done by you during this period is just the sum of $F(s)ds$ from the point where it starts until the end. This sum is just the area under the curve. So in general when we're dealing with a force that varies while the object travels, the work done is just the area under the force vs. distance curve.

Let's work out the details for a force we know varies with distance; the Hooke's law force exerted when you stretch an elastic object. Picture a spring with spring constant k , and imagine stretching it from its equilibrium position until it is stretched by an amount Δx . What is the work done on the spring in this process? First, let's make a picture of the force vs. distance curve for this:



To stretch the spring, you must apply a force along the direction of motion, so you do positive work when stretching it. How much work must you do to stretch such a spring a distance Δx ?

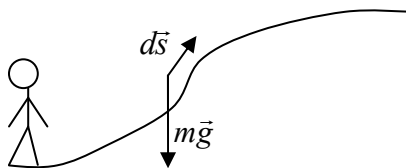
$$\text{Work} = \int \vec{F} \cdot d\vec{s} = \int_0^{\Delta x} kx dx = \frac{1}{2} k (\Delta x)^2$$

Stretching a spring takes work. To do this, you have to pull along the direction of motion. But something is different here. In the cases we have seen before, the work done on an object changed its kinetic energy. In stretching a spring, you are definitely doing work, yet the kinetic energy of the spring doesn't change. What does change? Where does the energy associated with that work go?

As we will see in the next chapter, the energy given to the spring by the work you do is not lost, energy never is. Instead of appearing as kinetic energy, it appears as 'elastic potential energy'; it gets stored in the stretching of the spring. Stretch the spring and your work stores energy in it. Let it go and that energy will come back out again, this time in the form of kinetic energy in the spring.

Work and objects traveling in curved paths

In the same way, we can calculate the work done by a force acting on an object traveling in a curved path. Consider walking up a hill:



As you go up this hill, gravity always acts straight down. For each little part of the path $d\vec{s}$ which is directed at an angle θ away from the downward direction of the weight, the work done by gravity is:

$$dW = m\vec{g} \cdot d\vec{s} = mgds \cos(\theta) = -mgdh$$

In this derivation, we have used the fact that $ds \cos(\theta) = -dh$, just a measure of how much higher the object moves during this motion. So the total work done by gravity added up along any path is:

$$W_{\text{gravity}} = \int \vec{F}_{\text{gravity}} \cdot d\vec{s} = -mg\Delta h$$

The work done on a moving object by the gravitational force depends on the magnitude of its weight and the amount by which the object's height changes. Notice in particular that it doesn't depend on the path the object takes. All that matters is where it starts and where it stops. We will see in the next chapter that this path independence makes the work done by gravity easy to account for.

How rapidly work is being done: power

Power is the rate at which work is done, or more generally, the rate at which the energy of a system changes. For some purposes, this is a useful quantity to keep track of. Since work represents a change in energy, the rate at which work is done is also the rate at which energy changes.

$$\text{Power} = P = \frac{dW}{dt} \quad \left(\text{or } \frac{dE}{dt} \right)$$

The units for power are Joules/sec. This unit, 1 J/s, is called a Watt (after James Watt, a Scottish inventor who made great improvements to the steam engine).

When we consider the work done by a *constant force*, we can write the power in a special form by recalling that the work done is given by $dW = \vec{F} \cdot d\vec{s}$, so that

$$P = \frac{dW}{dt} = \frac{d(\vec{F} \cdot d\vec{s})}{dt} = \vec{F} \cdot \frac{d\vec{s}}{dt} = \vec{F} \cdot \vec{v}$$

So in this special case in which a constant force acts, the power generated by it is the dot product of the force vector with the velocity vector.

Work and energy in a typical trip

In Chapter 7 we considered a kind of standard trip from the perspective of force and momentum. Let's consider this again in the context of work and energy. Let's take as our example a simple trip you might take on your bike. As you start out, you pedal so that your tires push backward on the ground. The ground, in turn, pushes forward on you. It is this forward force which pushes you and your bike down the street. This force acts along your direction of motion, and does positive work on you. Initially, it's the only force doing work, and your kinetic energy increases in response.

Once you start moving however, a second force begins to act, fluid friction. As you speed up, this force becomes larger, until eventually it balances the forward force from your peddling. The forward force continues to do positive work, putting energy into you. But now the friction force, which acts opposite your motion, does negative work at the same rate. The energy put in by the forward force is immediately drained away by the friction force; the positive work of one balanced by negative work of the other.

Eventually you stop peddling and glide to a halt. Once you stop peddling, the forward force disappears, while the friction force remains. It does negative work, draining away your kinetic energy. The magnitude of this fluid friction decreases as you slow, eventually dropping to zero as you come to a stop. What happens if you brake instead of gliding to a halt? In this case a suddenly large friction force acts, quickly draining away the energy you built up on acceleration.

If you keep the forward force constant, the power required increases as you speed up. You can see this in two ways. First, the distance traveled in each unit of time increases as you speed up. Work is force multiplied by distance, so the work done in each unit of time increases as you speed up. A second way to see this is to refer to the other definition of the power extracted when a force acts: $P = \vec{F} \cdot \vec{v}$. In this form it's clear that the power expended increases as you speed up.

In fact, most systems which might propel you, from the engine of your car to your legs when you run, are much closer to constant power than constant force systems. As such, they maintain the product of force times velocity, initially supplying a large force at small velocity, and eventually providing a small force at large velocity. The various gears in your car are designed to allow it to provide you with large force at a wider range of velocities.

9.5 Rotational energy: torque, angular displacement, and work

There is another way in which the energy of motion appears, and while it is fundamentally just the same old kinetic energy we have explored for point objects, it is useful to account for it in a slightly different way. Imagine an extended object, like a dinner plate. Such a plate may move in various ways. It might fly across the room with a large translational motion. It might remain in place, but spin about its center with a large rotational motion. Or more generally, it might both translate and rotate; spinning while it flies across the room.

The total energy of motion in the plate will always be the kinetic energy of each of the bits which make it up. But it is often useful to split this into a translational part and a rotational part. Let's begin with a simple case, a rotating disk. Each piece of the disk has a mass dm and a velocity v , and hence a kinetic energy. For a disk rotating with an angular velocity ω , each little part has a velocity $v = r\omega$. This makes the kinetic energy of each little piece $dKE = \frac{1}{2}dmr^2\omega^2$. Breaking the disk into concentric rings, each with a mass $m(r) = 2\pi r\rho dr$, we can write the total kinetic energy as:

$$KE_{\text{rotating disk}} = \int_0^R 2\pi r \rho \left(\frac{1}{2} r^2 \omega^2 \right) dr = \pi \rho \omega^2 \int_0^R r^3 dr = \frac{1}{4} \pi \rho \omega^2 R^4 = \frac{1}{4} MR^2 \omega^2$$

Remembering that the rotational inertia of such a disk is $I_{\text{disk}} = \frac{1}{2} MR^2$, we can rewrite this as:

$$KE_{\text{rotating disk}} = \frac{1}{2} I_{\text{disk}} \omega^2$$

Notice how completely parallel this is to linear motion. For linear motion, kinetic energy is given by one half times the inertia times the velocity squared. For rotational motion, kinetic energy is given by one half times the rotational inertia times the angular velocity squared.

How does rotational kinetic energy change? Once again, analogy gives a clue. Rotational kinetic energy is changed by rotational work.

$$W_{\text{rotation}} = \int \vec{\tau} \cdot d\theta = \Delta KE_{\text{rotational}}$$

When a torque acts while an object rotates through some angular displacement, rotational work is done. When a force acts while an object moves through some distance, ordinary work is done. In both cases, the total work done on a system is equal to the change in kinetic energy of the system.

9.6 Quantifying energy in technology and life: some comparisons

Energy makes everything happen, including life. It is essential for humanity. Ingested as food it enables our bodies to work the machinery of life. We bring energy into our homes to maintain warm surroundings in the winter, and expel it to keep things cool in summer. We invest energy in ourselves and our products to get them moving, and remove it again to bring them to a stop. Nothing that happens occurs without the transfer of energy. As a result, energy is a commodity we regularly pay for.

Oddly, we don't pay for energy in its natural units. You can't go to the gas station and buy a Joule of gasoline or of candy bars. We pay instead for quantities of the materials which store energy, a gallon of gasoline or a Milky Way bar. When you purchase electricity you do pay for energy, but it is usually not accounted in Joules. Most of it is measured in a funny mixed unit called a "kilowatt-hour". For one kilowatt-hour of electrical energy, you might pay 9 cents.

What is this unit really? One kilo-watt hour is 1000 Watts * 3600 seconds = 3.6×10^6 Joules. That's right, 3.6 million Joules; an awful lot of energy. Perhaps a comparison will help put it in context. If you weigh 80 kg, you'd have to be moving at 300 m/s to have this much kinetic energy. That's about 675 mph, a healthy (or maybe unhealthy) speed indeed! If someone said they'd speed you up to 675 mph for 9 cents, you'd probably think it pretty cheap.

Right now, people in the US use, on average, about 8000 kilowatt-hours per year. That's about 22 kilowatt-hours per day. The cost at current rates is a few dollars a day, and hence not prohibitive.

Another comparison might be helpful. To stay alive, you have to eat regularly. A typical adult diet might contain 2000 Calories a day. The "Calorie" of nutrition is equal to about 4184 Joules, so the 2000 Calories a day you need provides about 8.4×10^6 Joules of energy. What would this daily consumption of energy be in the electrical units we pay for, in kilowatt-hours? The 2000 Calories a day you must consume is about equal to two kilowatt-hours. If you paid as much for your food energy as you do your electrical energy, your daily meals would add up to about 20 cents a day.

You can see that the price we pay for electrical energy is very low compared to what we pay for food energy. There are a lot of reasons for this, but it's worth noting that electrical energy is largely extracted from the long ago stored resources of fossil fuels, while your food comes more directly from the annual collection of sunlight by plants and the animals that eat them. In this very real sense, food is a renewable energy resource. The higher cost of your daily 2000 Calories of food has many origins, but a good part of it is due to the fact that it is fundamentally renewable. The food you eat each year relies largely on the energy which passed through the Earth this year.

It is useful to remember some simple conversions that help you to place power and energy in a context:

The power obtained burning one gallon of gas in an hour = 39,000 watts

The power to run a light bulb = < 100 watts

So if you want to save energy, don't bother to turn off the lights, just stop driving your car. When you drive, you use energy at a rate hundreds of times larger than you do when you light a lamp.

The table below gives an idea of the relative cost of various energy sources. This is changing fast, but the basic pattern, that "renewable" food is 30-40x as expensive as our more familiar "energy sources" is pretty robust. Now food is not the cheapest way to harvest renewable energy. Among other things, we pay a lot to keep it tasty. Electricity generated from renewable resources is currently only a few times more expensive than electricity from the coal and nuclear sources we use now.

It may shock you to see how much those batteries really cost, Joule for Joule...

Type	Unit	Cost/Unit	Cost per Mega-Joule	Uses
Electricity	1 Kwh	\$0.10	\$0.027	appliances, motors
Gasoline	1 gallon	4.20	\$0.032	transportation
Natural Gas	1 Therm	0.60	\$0.011	Heating
Milky Way candy bar	1 bar, 1.9 oz	0.90	\$0.94	Food
AA battery	1 battery	0.40	\$80.00	portable electronics

It may also be useful to put our collective consumption of energy into these units. The annual usage of energy in the United States was about 10^{20} Joules in 2005. One 'standard' 9000 ton trainload of coal would have about 100 coal cars, each carrying 90 tons. Each car is about 53 feet long, so this train stretches a full mile. This mile-long train contains about 3×10^{14} Joules of energy. Sounds like a lot, but if all the US energy supply derived from coal, we would need 300,000 mile-long trains of coal to supply it. In fact quite a lot of our energy comes from oil and gas as well. So we use 'only' about 150,000 mile-long trains of coal a year. Along with all this coal, residents of the United States consume about 7.3 billion barrels of oil a year.

Now you might think these numbers are large just because the population of the United States is large, currently about 307 million. So let's look at per capita numbers. Each year, every person uses an average of 4.4 tons of coal, and 24 barrels of oil. Imagine what it would be like if you had to stack that all around your house...

A Quick Summary of Some Important Relations

Work creates changes in kinetic energy:

When forces act while an object moves, they may do work on it, changing its kinetic energy.

$$\text{Work} = \int_{\vec{s}_i}^{\vec{s}_f} \vec{F} \cdot d\vec{s} \quad \text{KE} = \frac{1}{2}mv^2 \quad \int_{\vec{s}_i}^{\vec{s}_f} \vec{F} \cdot d\vec{s} = \Delta\text{KE}$$

Remember that because of the dot product in the definition of work, only forces which have some component along (or opposite) the direction of motion do work. If they are along the motion, they increase kinetic energy. If they are opposite the motion, they decrease kinetic energy.

Work done by gravity:

Gravity does work only when objects move up or down, and the amount of work done depends only on how much they move up or down.

$$W_{\text{gravity}} = -mg(y_f - y_i) = -mg\Delta h$$

Power and the rate of energy change:

Power measures the rate at which work is done (or the rate at which energy changes). When a constant force is acting, power is the product of force and velocity.

$$P = \frac{dW}{dt} = \frac{d(\text{KE})}{dt} \quad P_{\text{constant F}} = \vec{F} \cdot \vec{v}$$

10. What could happen: the work of conservative forces and potential energy

- 1) The work done by gravity
 - i. Path independence and reversibility
 - ii. Accounting for gravity's work: potential energy of position
 - iii. When only gravity does work: changes in energy depend on position alone
 - iv. Constraining forces and work: roller coasters and swinging on swings
- 2) Gravitational potential energy and its applications
 - i. The pendulum
 - ii. Sliding or rolling without friction
 - iii. Where does potential energy reside?
- 3) Elastic potential energy and deformation
 - i. Linear materials
 - ii. Non-linear material: the J-curve
 - iii. Stored energy, fracture, and toughness
- 4) Potential energy curves and motion
- 5) Potential energy and conservative forces

Physics for the Life Sciences: Chapter 10

We have seen in the last chapter that work, the action of a force through a distance, alters the kinetic energy of an object. In this chapter we will see how, for some forces, it makes sense to account for the work which they *could potentially* do as another form of energy. We will analyze the details of this notion for two quite familiar cases.

The first is gravity. When you stand on a rooftop, it is clear that your state is tangibly different from when you stand on the street. Looking out over the edge, you are quite aware of what gravity could potentially do to you. Take one false step and you know it would start doing work on you, increasing your kinetic energy rapidly until you reached the ground below. The work which gravity is prepared to do on you is a kind of energy it is ready to supply. We can usefully keep track of how much energy gravity is ready to provide by talking about 'potential energy'.

A second example involves elastic forces. Perhaps during your childhood you 'shot' a few rubber bands at your classmates. In doing this, you stretched out the rubber band, then suddenly released it. That stretched band quivered with vitality, you could sense that it was ready to release the energy you had put into it and fly off toward its target. It too was ready to act, it had the potential to do work and generate kinetic energy.



In what follows we will learn how to quantify potential energy. We will use as examples the potential energy associated with gravity near the Earth's surface and with the elastic deformation of materials. Understanding potential energy is the next important step in enriching our understanding of energy and how to keep track of its flow in the world. In later chapters we will include more directly the potential energy associated with stretching the individual bonds between atoms, something remarkably analogous to stretching elastic materials.

10.1 Work done by gravity and potential energy

Last time when we discussed work done on an object moving in a curved path we looked at the specific case of work done by the force of gravity as an object moves from one place to another. In this case we found:

$$W_{\text{gravity}} = \int \vec{F}_{\text{gravity}} \cdot d\vec{s} = -mg(h_f - h_i) = -mg\Delta h$$

When we derived this, we noted an odd feature that it has. The work done by gravity as you go from some initial location to some final location depends *only* on where you start and stop, and not at all on how you get there. Another, fancier way to describe this is to say that the work done by gravity when moving from one place to another is "path independent".

We need to think about what this means. Imagine an object traveling *only under the influence of gravity*, something like a projectile flying without friction. As it moves, gravity does work on it, changing its kinetic energy:

$$W_{\text{gravity on projectile}} = -mg\Delta h = \Delta KE_{\text{projectile}}$$

Since the work done by gravity depends only on position and *not* on path, the final kinetic energy of this object is determined completely by where it is:

$$\begin{aligned} \Delta KE &= KE_{\text{final}} - KE_{\text{initial}} = -mg\Delta h \\ KE_{\text{final}} &= KE_{\text{initial}} - mg\Delta h \end{aligned}$$

Let's repeat that: for an object acted on only by gravity, changes in kinetic energy are simply a function of position. You tell me where an object started out and where it is now, and I'll tell you what its kinetic energy is. Consider a simple example: if you toss a ball into the air it begins with large kinetic energy. As it rises, gravity does negative work on it, taking away some of its kinetic energy. Then as it falls, gravity does positive work on it, giving that energy back.

Gravity has an unusual property. Any energy it takes away from an object as it rises can be returned to it as it falls. Since it can always restore any energy its work takes away, it is useful to think of the work done by gravity as "storing up" some of the energy of the ball, keeping it fully available for eventual return. This kind of stored energy is something we call **potential energy**. Imagine that you throw the ball straight up. Its kinetic energy is converted to potential energy as it rises. At the top there is a moment when all of its kinetic energy has been converted to potential energy, and then as it falls the potential energy is turned back into kinetic energy.

How shall we measure this potential energy? It turns out the appropriate definition is:

$$\Delta PE_{\text{gravity}} = -W_{\text{gravity}}$$

Pay close attention to the minus sign in this definition. It's really crucial. It says that when gravity does positive work on the system, when gravity increases the kinetic energy of the object, the potential energy *decreases*. This is as it should be. To increase the kinetic energy, some of the energy stored as potential energy must be used up. By contrast, when gravity does negative work on the system, it reduces the kinetic energy of the object, then the potential energy of the system increases. Again, this makes sense. Now gravity is taking away kinetic energy and storing it as potential energy.

For gravity, changes in potential energy are directly related to changes in position:

$$\Delta PE_{\text{gravity}} = -W_{\text{gravity}} = mg\Delta h$$

As the height of an object changes, its gravitational potential energy changes. If it goes higher (Δh positive), the potential energy is becoming larger. If it goes lower (Δh negative), the potential energy is becoming smaller. Notice too that we're not defining any *absolute* potential energy. This definition only allows us to track *changes* in the potential energy. That's consistent with our overall understanding that physics cares only about changes in energy, and not its absolute value.

Energy when only gravity acts

Now *if no force but gravity does work*, we can use the work-kinetic energy theorem to notice something interesting:

$$\Delta KE = \sum W_{\text{all forces}} = W_{\text{gravity}}$$

or, moving the work to the other side and using the definition of potential energy:

$$\Delta KE - W_{\text{gravity}} = 0 \quad \text{or} \quad \Delta KE + \Delta PE_{\text{gravity}} = 0$$

As long as no other forces do work, any change in kinetic energy is just balanced by a change in gravitational potential energy. We can freely transfer the energy back and forth between potential and kinetic forms, but any increase in kinetic energy must be balanced by a decrease in potential energy, and vice versa.

There is another instructive way of writing this:

$$\Delta KE + \Delta PE_{\text{gravity}} = KE_f - KE_i + PE_{\text{gravity}}^f - PE_{\text{gravity}}^i$$

OR

$$KE_f + PE_{\text{gravity}}^f = KE_i + PE_{\text{gravity}}^i$$

In this form you can clearly see that this is a conservation law. There is some quantity, the sum of kinetic and potential energy. Whatever this total is at the initial time, the total of the two remains exactly the same at any later final time. This quantity, the total of kinetic and gravitational potential energy, is **conserved**; it never changes with time.

Remember though, to apply the work-kinetic energy theorem as we did here, we need to be sure that gravity is the *only* force doing work on the object. If air friction, for example, is present, it too would do work on the object. Then we would write:

$$\Delta KE = \sum W_{\text{all forces}} = W_{\text{gravity}} + W_{\text{friction}}$$

And

$$\Delta KE - W_{\text{gravity}} = \Delta KE + \Delta PE_{\text{gravity}} = W_{\text{friction}}$$

$$KE_i + PE_{\text{gravity}}^i = KE_f + PE_{\text{gravity}}^f + W_{\text{friction}}$$

What would this mean? In this case, instead of just converting potential energy into kinetic energy, some of it is lost. We account for how much by calculating the work done by friction. What happens to this energy? When friction does work, where does the energy go? The work done by friction describes the transfer of energy to another form, different from kinetic or potential energy. Friction converts these forms of energy into 'thermal energy'. We will explore the nature of this thermal energy in some detail in subsequent chapters.

Constraining forces and energy

It is often the case that an object will be constrained to move on a particular path by a force. A good example of this is a pendulum. The tension which keeps the pendulum moving in a circle always acts in a direction perpendicular to the direction of motion. As a result, it never does work on the object. Since the tension never does work, the only forces which do work while the pendulum swings are gravity and air friction. We'll ignore the friction for a start, and imagine what would happen if gravity was the only force doing work on the pendulum. Such a pendulum swings back and forth, trading energy between gravitational potential energy at the top and kinetic energy at the bottom. The other force which acts, the tension, does no work at all, and hence never alters the energy.

Likewise, if I take a ball and roll it inside a bowl, the force which keeps it from moving through the bowl is the normal force. Again this normal force always acts perpendicular to the motion, so it can never do work, it can never change the energy of the ball. So again, it just rolls back and forth, trading energy between kinetic and potential, and back again.

Here are some other cases where the same sort of thing happens:

1. Roller-coasters (the very name "coaster" evokes this)
2. A ball rolling down a hill
3. A child swinging on a swing
4. A bicyclist coasting up or down a hill

We have left something obvious out of course. In each of these cases, there is a force other than gravity and the force of constraint acting: friction. In these cases, friction always acts opposite the direction of motion. Because of this, it always does negative work. This negative work gradually drains away the energy which would otherwise be trading back and forth forever between kinetic and potential. Eventually, the negative work done by friction ‘drains away’ all the energy, and the object comes to rest, sitting at the lowest point.

10.2 Gravitational potential energy relative to a reference location

We have said that we can ‘store’ energy in the work done by gravity and shown how it’s useful to talk about this as a form of energy. Now we’d like to say how much potential energy an object has. But it’s impossible to say what the ‘absolute’ potential energy of anything is. Just as the absolute value of kinetic energy is observer dependent, so too the absolute value of potential energy is meaningless. Remember how we defined potential energy:

$$\Delta PE_{\text{gravity}} = -W_{\text{gravity}}$$

We never defined the absolute potential energy, but only changes in it. So it is meaningless to talk about the absolute potential energy. It is only useful to speak of the change in potential energy as an object moves from one location to another.

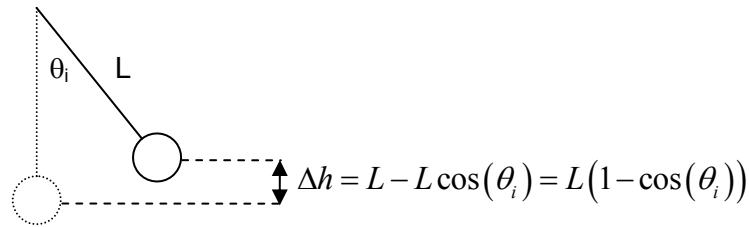
On the other hand, it is sometimes convenient to measure the potential energy of an arrangement *relative* to some reference point. For example, we might define the potential energy of a projectile to be zero when it is resting on the ground. Since the absolute value of the potential energy is always arbitrary, there’s no reason why we can’t do this. If we do, then we could calculate the change in potential energy moving from the ground to any particular point and just call that ‘the potential energy’ of the object. In this case, we might write:

$$PE_{\text{gravity}} = mgh_{\text{above ground}}$$

This approach is quite commonly taken when talking about potential energy, and can be very convenient. But when we do this, we must always remember that it has meaning *only* if we defined the potential energy to be zero when the object is on the ground. If we do this, any object above the ground has positive potential energy (relative to what it would have at the ground), and any object below the ground has negative potential energy (again, relative to what it would have at the ground).

An example: the velocity of a pendulum

Consider the motion of a pendulum. We will release it from rest at some angle θ_i from the vertical:



What will its velocity be at the bottom? To answer this question using Newton's laws is difficult, because the tension T is constantly changing as the pendulum moves. This question is quite easy to answer from the energy approach however. We know from the work-energy theorem that:

$$\Delta KE = \sum W_{\text{all forces}} = W_{\text{tension}} + W_{\text{gravity}} + W_{\text{friction}}$$

The tension does no work because it is a force of constraint (it always acts perpendicular to the motion). If the work done by friction is small enough to ignore, then:

$$\begin{aligned} \Delta KE = W_{\text{gravity}} &= -\Delta PE_{\text{gravity}} & \Delta KE + \Delta PE_{\text{gravity}} &= 0 \\ KE_f - KE_i + PE_f - PE_i &= 0 \end{aligned}$$

Let's call the point at the bottom the final position and, as we suggested above, define the potential energy at the bottom of the pendulum's swing to be zero. If we do this, we can rewrite the above as:

$$\frac{1}{2}mv_f^2 - 0 + 0 - mgL(1 - \cos(\theta_0)) \quad \text{or} \quad v_f = \sqrt{2gL(1 - \cos(\theta))}$$

What if I ask for the velocity of the pendulum at some other angle, call it θ_1 ?

$$\frac{1}{2}mv_1^2 - 0 + mgL(1 - \cos(\theta_1)) - mgL(1 - \cos(\theta_0)) \quad \text{or} \quad v_1 = \sqrt{2gL(\cos(\theta_1) - \cos(\theta_0))}$$

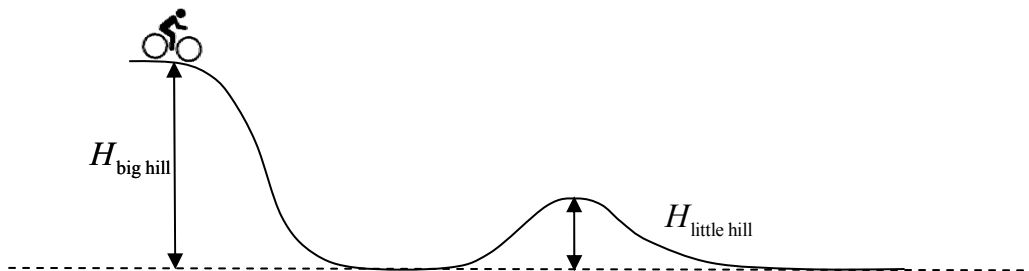
It is often the case that a very 'hard' problem in the force picture can become very easy in the energy/momentum picture. This is a fine example. In a sense this is because the energy approach concentrates on what's really important in the problem, ignoring things (like forces of constraint) which don't play a crucial role.

A second example: sliding or rolling without friction

Another nice example is the motion of an object which moves on a hilly surface, pulled down slopes by gravity and carried up them by momentum. To be concrete, let's consider a bicyclist who starts at the top of a hill, coasts to the bottom, then continues up the next. Just as in the pendulum problem, we track changes in kinetic energy of the cyclist by accounting for the work done on him.

$$\Delta KE = \sum W_{\text{all forces}} = W_{\text{gravity}} + W_{\text{friction}} + W_{\text{normal force}}$$

Since the normal force is always perpendicular to the direction of motion, it never does any work. We will assume, for a start at least, that there is no friction, either opposite or along his motion. Under these assumptions, what's the velocity of the cyclist at the bottom of the hill?



The only force which does work (in our simplified model) is gravity, and we know that potential energy accounts for the work it does. So we can write:

$$\Delta KE = KE_{\text{final}} - KE_{\text{initial}} = W_{\text{gravity}} = -mg(H_{\text{final}} - H_{\text{initial}})$$

$$\frac{1}{2} m_{\text{cyclist}} v_{\text{bottom}}^2 - 0 = m_{\text{cyclist}} g H_{\text{big hill}}$$

Or

$$v_{\text{bottom}} = \sqrt{2gH_{\text{big hill}}}$$

How fast would the cyclist be going when he reaches the top of the second little hill?

$$\frac{1}{2} m_{\text{cyclist}} v_{\text{little hill}}^2 = -mg(H_{\text{little hill}} - H_{\text{big hill}})$$

$$v_{\text{little hill}} = \sqrt{2g(H_{\text{big hill}} - H_{\text{little hill}})}$$

What if the cyclist is participating in the X-games, and wants to race down the first hill, zoom up the second hill, and actually go airborne, flying off the second hill instead of remaining on the ground? Could we figure out how high the first hill has to be for this to happen?

Picture the cyclist going over the second hill. We might model the hill as a circle with a radius equal to its height. In this model, for the cyclist to travel over the top, he must experience a centripetal force of known size:

$$F_{\text{centripetal}} = \frac{mv^2}{r}$$

Inserting what we know about the velocity of the cyclist here, and setting the radius equal to the height of the little hill, we get:

$$F_{\text{centripetal}} = \frac{m \left(2g \left(H_{\text{big hill}} - H_{\text{little hill}} \right) \right)}{H_{\text{little hill}}} = 2mg \left(\frac{H_{\text{big hill}}}{H_{\text{little hill}}} - 1 \right)$$

Where force is available to provide this centripetal force? The only force holding the cyclist on the ground is his weight. If the required centripetal force is *greater* than his weight, there will not be enough force to keep him on the circle. Using this equality, we can find the minimum $H_{\text{big hill}}$ which would allow the cyclist to jump off the second hill:

$$2mg \left(\frac{H_{\text{big hill}}}{H_{\text{little hill}}} - 1 \right) = mg$$

$$H_{\text{big hill}} = \frac{3}{2} H_{\text{little hill}}$$

Now that's an interesting, very simple result. To 'jump off' the second hill, just make sure the first is at least 1.5 times as big as the second. But remember, to get this result we built a really simplistic model, assuming that no friction acted at all. Surely this is not true. How would the result change if we relaxed this assumption?

First, imagine what happens if there is only friction which opposes the motion. For a cyclist, this would largely be air friction. It would always do negative work on the cyclist, reducing his kinetic energy. Since he would be losing energy in this way, the big hill would have to be larger than we calculated here if he was still to make his jump.

But that's not the only kind of friction he could experience. Imagine that, during his ride, he pedals forward hard. If he does, his tires will push backward on the ground, and friction with the ground will push forward on him. This *forward* frictional force will do positive work on the cyclist, adding to his kinetic energy. In this case he may arrive at the little hill traveling *faster* than he would be if only gravity acted. In this case, the big hill could be smaller than what we have calculated here. Indeed, if he pedaled hard enough, he could make the jump completely without the first hill.

Many variants of this kind of problem exist. In each case, we might start by ignoring the effects of friction and accounting only for the work done by gravity. Once we have those basic answers then we can begin to explore how the answers might change if other forces acted as well.

Where does the potential energy reside?

When we speak of kinetic energy, it's quite clear where the energy resides; in the object which is moving. What about potential energy? When you throw a ball upward, it begins with substantial kinetic energy. As it rises, the force of gravity does negative work on it, extracting the kinetic energy from it. We describe the kinetic energy taken from the ball as moved to gravitational potential energy, knowing that this energy might be returned to the ball.

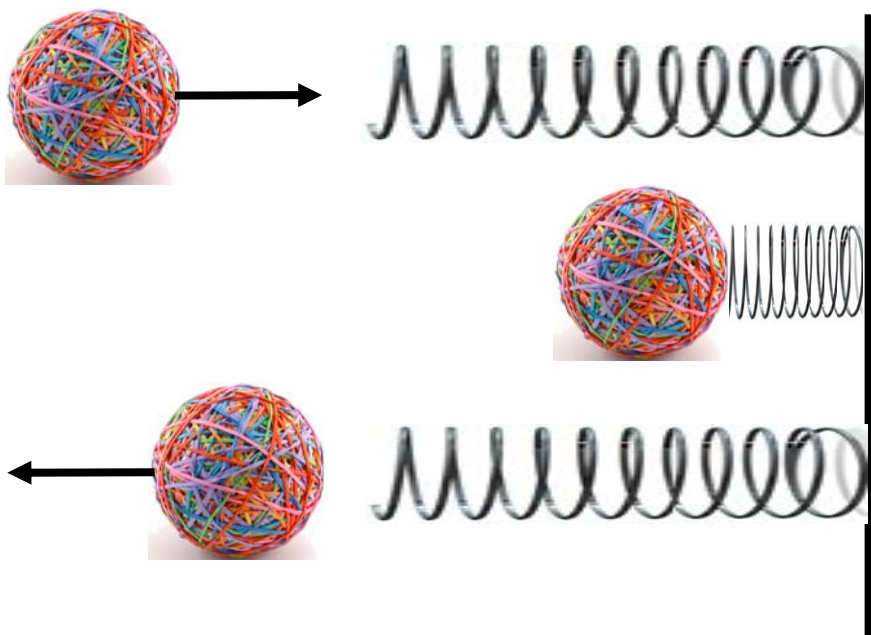
Where is this potential energy? Is it in the ball or the Earth? In fact the gravitational potential energy is stored in the particular arrangement of the Earth and the ball; it arises because of their gravitational interaction. When the Earth and the ball are farther apart, their potential energy is larger. When they close

together it is smaller. But in any case, both objects must be present for the potential energy to exist. Potential energy is not stored in the ball, or in the Earth, but in a particular arrangement of *both* the ball and the Earth, separated in a certain way. The potential energy is really a property of the system, rather than of either object.

10.3 Elastic potential energy of a deformed material

Let's consider briefly another way of storing energy, another form of potential energy. If I fire a ball at a spring it will arrive at the spring with some kinetic energy. The spring will gradually slow it down until it has no kinetic energy; the ball comes to a stop. Now the spring is compressed, and continues to push back against the ball, accelerating it back in the opposite direction from which it arrived, until eventually it leaves the spring with just the energy it arrived with.

The spring absorbs the energy of the moving ball, and then restores it to the ball as it flies off. Note the similarity to gravity. This is just what happens when you throw a ball up in the air; its kinetic energy is taken away and stored as potential energy, then return to it as it falls. Let's see how to quantify this for the spring using the work-energy theorem.



We will assume the spring deforms linearly, and model the force applied to the ball using Hooke's Law.

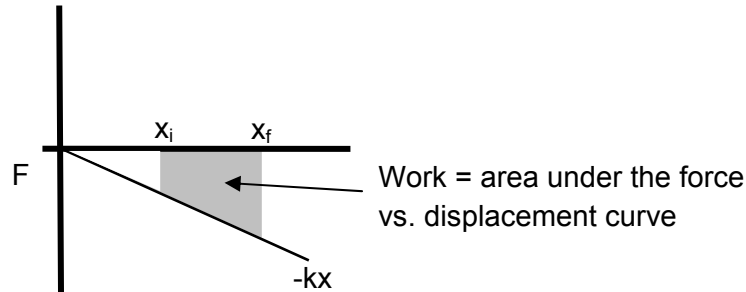
$$F_{\text{spring on ball}} = -k\Delta x$$

In this model Δx measures how much the ball has compressed the spring, while k is the 'spring constant'; a number which describes how stiff the spring is. For a stiff spring, k is large, for a smushy spring k is small. Remember that the minus sign means the force is opposite the direction of the deformation " Δx ". So when the ball compresses the spring in the positive x direction, the force is in the negative x direction.

How much work does the spring do on the ball?

$$W_{\text{spring on ball}} = \int_{x_i}^{x_f} \vec{F} \cdot d\vec{s}$$

Let's draw a picture of the magnitude of the force exerted by a spring at each point. We start with $x = 0$ at the point where the spring is un-stretched.



The work done by the spring at each point is the force at that point $F(x)$ multiplied by the displacement dx . Adding this up as the spring is stretched from $x = x_i$ to $x = x_f$ amounts to calculating the area under the curve in the figure shown above. This total area, which is negative because the force is in the direction opposite the displacement, is:

$$W_{\text{spring on ball}} = -\frac{1}{2} kx^2 \Big|_{x_i}^{x_f} = -\left(\frac{1}{2} kx_f^2 - \frac{1}{2} kx_i^2\right)$$

Note the minus sign. This means that as the ball compresses the spring, the force is always opposite the motion, always doing negative work on it.

How does this relate to the storage of energy in the spring? What is the 'potential energy' stored in the spring? We know by definition that:

$$\Delta PE_{\text{elastic}} = -W_{\text{spring}}$$

so for a spring:

$$\Delta PE_{\text{spring}} = -W_{\text{spring}} = \frac{1}{2} kx_f^2 - \frac{1}{2} kx_i^2$$

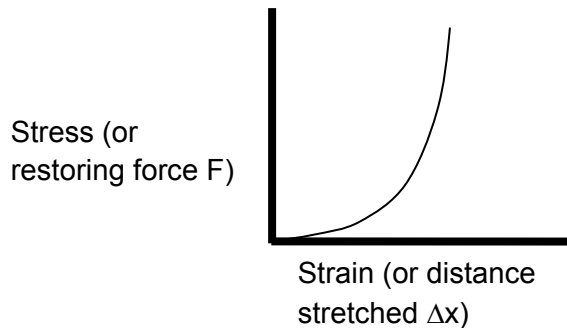
Now this is an interesting equation. For a start, imagine what happens if $x_i = 0$. In this case, the spring starts out unstretched, with no particular energy stored in tension or compression. It might make sense to define the potential energy stored in the unstretched spring to be zero. If we do this, the equation above becomes simply:

$$PE_{\text{spring relative to unstretched}} = \frac{1}{2} kx^2$$

The potential energy stored in such a spring is always positive, whether we stretch or compress it. Any time I either stretch or squash a spring, I store energy which the spring can return to me by relaxing to its unstretched length. Returning to the start of this chapter, you can see that the stretched rubber band, ready to be shot off across your classroom, is in fact storing this kind of energy in its stretched state.

Non-linear materials and stored energy

Earlier in this book, way back in Chapter 4, we noted that many biological materials deform under stress in ways which differ from the linear (Hooke's Law) response analyzed in the last section. How does this affect the energy stored in the deformation? The basic principle is exactly the same, with the energy stored in the deformation being equal to the work done to stretch the material. Let's start by recalling the nature of the typical strain vs. stress curve for such a material:



We called this the “J-curve” because of its shape. This curve suggests that while you can stretch the material a lot by applying just a small force, if you want to continue to stretch it further, you'll have to apply a larger and larger force. Remember that this stress-strain curve represents the force you would have to apply along the direction of motion to achieve the desired distortion. You do positive work on the material to distort it, storing energy in it. It does negative work on you, taking energy away and storing it in its distortion. How much energy is stored in this kind of deformation?

Just like in the spring case, we find this energy by determining the work done by the material as we stretch it. Remember, we want to find the work done by the material on the object which is deforming it. When an object pushes in on such a deformable material, the material pushes back. The work done on you by the material while you compress it is negative, just as the work done by the spring on the ball in the last section was negative. The material will do negative work, meaning the potential energy associated with this will be positive.

But notice how this differs from the Hooke's law case. With this J-curve shape, you can deform a material quite a lot (to large Δx) while the force F remains small. The area under the curve in the beginning is tiny, and the energy stored in the material while getting to a pretty large distortion remains small. Eventually of course, the J-curve starts to rise, and as you continue to distort the material, more and more energy will be required for each increase in distortion.

Why is this important for life?

Stored energy, fracture, and toughness

All of us have broken things: smashed a water glass, snapped a piece of string, knocked an icicle off the gutter. You may have noticed that when things break, it seems to happen “by itself”. When you throw a ball into a window, you don’t just punch a ball-sized hole in it; the whole thing shatters. How do things break, and how does this relate to the energy stored in deformation we’re just talking about?

To break a solid object you have to pull apart atoms which begin bonded together. Doing this requires energy, and that energy has to come from somewhere. As an example, here’s what happens when you drop a water glass you’re holding in your hand. It begins with a relatively large gravitational potential energy, relative to what it would have at the floor. Since nothing supports it, it accelerates toward the floor, converting that gravitational potential energy into kinetic energy.

When the glass hits the floor, the floor pushes up on it with a large force, meanwhile the glass pushes down on the floor with an equal and opposite force. These forces deform both the glass and the floor, squashing them just as an elastic ball deforms when it strikes the ground. Deforming the glass stores energy in it! That stored energy can be used to break the bonds between atoms in the glass. If enough energy is stored, there will be enough to send cracks propagating through the whole thing and it will shatter.

How about other things, like snapping the string or breaking the icicle? In these cases too, breakage occurs after you store energy in the material by deforming it. You stretch the string, or push the bottom of the icicle to the side. The energy you store by doing this is then released when the object breaks.

If you think a bit, you’ll realize that biological materials are a lot tougher than some of these manmade or crystalline materials. When you trip and fall you may bruise, but fortunately you don’t shatter. The J-curve is one of the important reasons for this. Many biological materials can stretch a *lot* while storing only a small amount of energy. This allows your tissue to deform smoothly while spreading the applied force around, avoiding the concentrations of stress which occur in stiffer materials.



10.4 Potential energy curves and motion

When we first defined the potential energy of a one dimensional arrangement, we wrote:

$$\Delta PE = -\int F ds$$

or

$$dPE = -F ds$$

from this, you should be able to see that anywhere the potential energy doesn’t change as you move around, the force must be zero. Anywhere the potential energy changes a lot when you travel a short distance, the force must be large. In fact, we can rewrite this as:

$$F = -\frac{dPE}{ds}$$

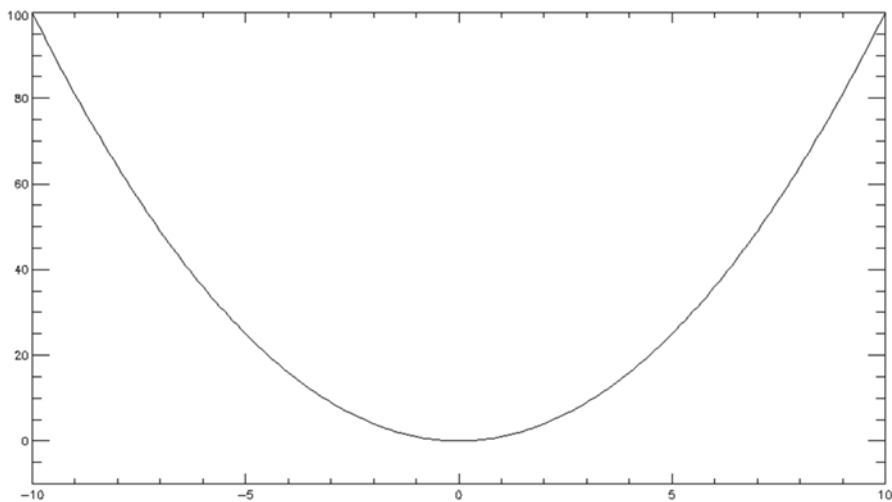
That is, the force at any point is given by minus the slope of the potential energy at that point. If we now represent the potential energy as a function of position by graphing it, the force is always minus the slope of the potential energy vs. position curve. Let's look at a simple example: the potential energy associated with a spring. Defining the potential energy of the unstretched spring to be zero, we can write:

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

Taking the derivative, we find, as we expect, that the force associated with this is:

$$F(x) = -\frac{dPE(x)}{dx} = -kx$$

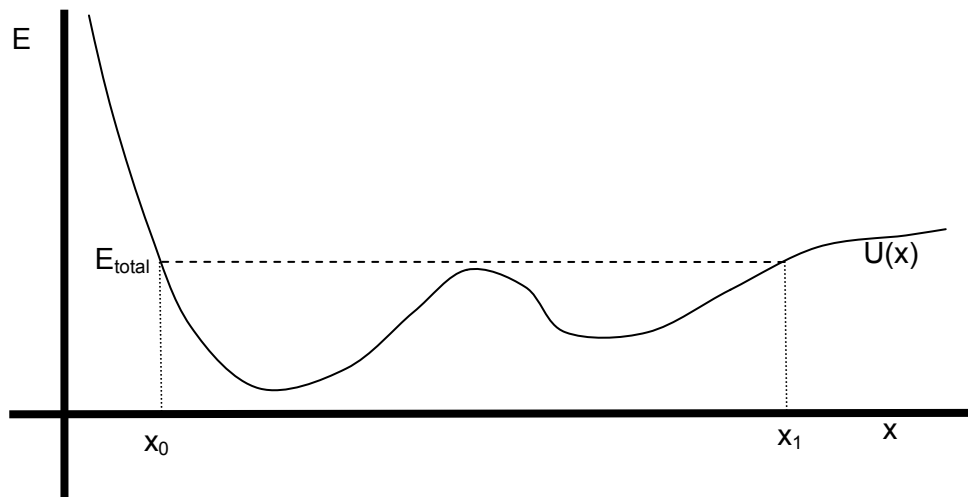
What does the graph of this potential energy function look like? It's very simple in this case, just a parabola, pointing upward, which has a minimum where $x = 0$. At this point, the minimum of the potential, the slope of the potential is zero, and hence the force is zero. This is the equilibrium point. Set the object here and it will happily stay forever. Set it at some other point, and it will experience a force pushing it back towards this.



Now imagine another potential energy function, something more complex. The very simple nature of the potential energy associated with gravity makes it easy for us to do this. Picture a landscape filled with hills and valleys. In this landscape the potential energy is largest at the top of the highest hill. It is smallest at the bottom of the deepest valley. On every slope, the potential energy changes as you move around. Everywhere the potential changes, there is a gravitational force which acts, and every time it will pull downhill.

Places where the slope of the potential goes to zero (the tops of hills and bottoms of valleys) are possible equilibrium positions. Places where the potential is minimized are positions of stable equilibrium. Move

away from such a minimum and, no matter what the details, you will be pushed back toward the equilibrium point. Every such potential minimum will be a stable equilibrium point, and any object moved a bit away from such a point will oscillate back and forth around it. Here's an example:



Imagine a cyclist moving on this landscape. At any point, the cyclist has some total energy $E_{\text{total}} = KE(x) + U(x)$. Let's imagine that E_{total} has the value shown as the dashed line on this figure, and that the cyclist begins at x_0 . At this point, her total energy is just equal to her potential energy. She has no kinetic energy at x_0 , and is not moving.

But this is not an equilibrium point. At this point the slope of the potential is negative, which implies a force in the positive x direction. So if she starts here, she'll be pushed to the right. As she moves right, her total energy will remain constant at E_{total} , but since the potential $U(x)$ is declining, her kinetic energy must increase. She will go faster and faster, always being pushed forward, until she reaches the minimum in the potential. Then, as she starts uphill, her potential energy will increase and her kinetic energy decrease.

As she goes over the hill, her kinetic energy will become quite small, just the small difference between the total energy and the potential energy at the top of the hill. But after she passes over the hill she will again gain kinetic energy. She will speed up until she passes through the next potential minimum. After this, she will again slow, more and more, until she finally comes to rest (though not equilibrium) at the point x_1 . After this, she'll be pushed back the other way, and will oscillate in this strange way back and forth between x_0 and x_1 .

What would she have to do to stop at the bottom of the first hill? As she moved from x_0 toward the bottom of the hill, she would speed up, trading potential for kinetic energy. To stop at the bottom, she would have to somehow eliminate that kinetic energy, so that when she got to the bottom of the hill her total energy would be just equal to her potential energy. She might do this by applying a non-conservative force like friction, rather than the conservative gravity which is otherwise pulling her down the hill.

10.5 Potential Energy and Conservative forces

This idea of potential energy is a very powerful one. It allows us to understand the work done (or which will be done) by some forces in a very simple way: to know the work done all we have to know is where things begin and end. Changes in energy are then just a function of position, and not of history.

Gravity is not the only force which we can define potential energy for. In fact we can define a meaningful potential energy for any force which is a "conservative" force. You can tell that a force is conservative if the work done by it when traveling between one location and another is independent of the path taken to get there. When this is true, the work done by the force is a function only of location, and a potential energy can be meaningfully defined.

There is an alternate way to state this requirement. If the total work done by a force as you travel in a closed path (starting and ending at the same point) is zero, the force is conservative.

The conservative forces we will encounter now include gravity and the elastic force associated with deforming a solid. These are both cases where the obvious macroscopic nature of the force obeys the requirements outlined above: work done by these forces is path independent. You will see later in your study of physics other forces which are conservative, like the electromagnetic force. They too will be associated with potential energies. In each case, these are forces which allow you to use the work done by the force to "store" energy which you can later completely recover.

But energy conservation of this kind is not at all obvious when we're talking about complicated, extended objects made up of many parts. For this reason we will sometimes speak of "non-conservative" forces. These non-conservative forces do work which is path *dependent*. As a result you cannot easily recover the energy these forces extract.

Imagine I slide a book from point A to B on a table. The force of kinetic friction as the book slides is always $\mu_k F_N$, and it always directly opposes the motion. So the work done by it is just:

$$W_{\text{friction}} = -\mu_k F_{\text{Normal}} \Delta s_{AB}$$

The amount of work done definitely depends on *which* path the book takes from point A to B. If I move the book from A to B, and then back from B to A, friction does negative work on it during both parts of the motion, always extracting energy from it.

Thinking about the second definition of a conservative force, you can see that the total work friction does around a closed path is *not* zero. Though I can take away energy from an object through the action of friction, I really can't get it back, at least not through the action of friction. So the work done by friction is not storing energy in a way which we can easily recover. It's not a conservative force.

Does this mean that energy is really not conserved? Is energy used up, or destroyed when friction acts? In fact energy, like momentum, is always conserved. When we speak of a non-conservative force in this sense, we're just saying that when this force acts, it converts the energy taken from the object into a form from which it is difficult to recover. Every bit of the energy removed from the object by a non-

conservative force still exists; it is just not easily returned to the object by the same force which took it away.

Another set of non-conservative forces are the active forces which you might apply to an object. When you push a chair across the floor, you do positive work on it, putting energy into it. Friction with the floor meanwhile does negative work on it, draining energy from it. At first, you put in more energy than friction drains, and the chair speeds up. Quite quickly, the energy you are putting in is balanced by the energy friction is removing, at which point the chair travels at a constant rate, and has a fixed energy. When you stop pushing, you stop putting energy in, friction continues to drain energy out, and the chair slows to a halt.

Notice that all these non-conservative forces (friction, the force you push something with, etc.) are described by *phenomenological* force laws, rather than fundamental ones. We will find, later in the class, that *all* of the fundamental forces are conservative. These ‘non-conservative’ forces we encounter are really nothing more than complex manifestations of underlying conservative forces.

A Quick Summary of Some Important Relations

Defining potential energy:

For 'conservative' forces like gravity, it is useful to account for the work they do as a way of storing potential energy. When the work done by such a force reduces KE, it 'stores' the energy as PE. When the work done increases KE, it pays for this by reducing PE. Only changes in potential energy are defined.

$$\Delta PE_{\text{cons. force}} = -W_{\text{cons. force}}$$

Gravitational potential energy:

$$\Delta PE_{\text{grav.}} = -W_{\text{grav.}} = mg\Delta h$$

Often we will account for potential energy by measuring changes relative to some reference point. If, for example, we define the PE at $y = 0$ to be zero, we could write:

$$PE_{\text{with PE(0) defined to be 0}} = mgy$$

Work, potential energy, and kinetic energy:

Changes in KE are determined by total work. If some of that work is accounted for as changes in PE, we can rewrite the work-energy theorem in a useful way.

$$W_{\text{nonconservative forces}} = \Delta KE + \Delta PE_{\text{conservative forces}}$$

This equation says that if nonconservative work is done (by friction, or a push), the object can lose or gain energy, increasing or decreasing the sum of kinetic and potential energy. If no nonconservative work is done, then the sum of kinetic and potential energies should be constant; it can change from one form to another, but the sum can never increase or decrease.

Elastic potential energy:

The potential energy associated with elastic deformation is defined by:

$$\Delta PE_{\text{elastic}} = \int_{x_i}^{x_f} -kx dx = -\frac{1}{2}k(x_f^2 - x_i^2)$$

We will often define the potential energy of the material when not stressed to be zero, and then write:

$$PE_{\text{elastic}} = \frac{1}{2}kx^2$$

11. Modeling isolated interactions: collisions

- 1) The basics: collisions as exemplars of interaction
- 2) Collisions between macroscopic bodies
 - i. A simple example: collision of two identical balls
 - ii. Elastic and inelastic collisions
 - iii. Collisions of unequal objects
 - iv. Collisions with non-contact forces like gravity
 - v. An inverse collision: two body decay and the discovery of the neutrino
 - vi. Another inverse collision: rocket propulsion
- 3) Collisions in three dimensions
- 4) Collisions of atoms and molecules
 - i. Can they store energy internally?
 - ii. The impact of quantization of internal energy states

Physics for the Life Sciences: Chapter 11

11.1 The basics: collisions as exemplars of interaction

In our discussion of energy so far, we have been talking about the kinetic and potential energy of whole objects. We saw that this bulk kinetic energy is a way of quantifying the overall motion of an object, either in translation or rotation. Likewise, we saw how the overall potential energy of an object was associated with a particular arrangement of things; the earth and a book, a ball and a spring. Both of these refer to the kinetic and potential energies of the whole object.

Our next major goal is to understand the kinetic and potential energy *inside* an object; stored in the motions and configurations of its atoms. All matter is made of atoms. Despite appearances, these atoms are *always* in motion; rattling around, banging into one another. In a liquid or a solid, their motion repeatedly stretches and relieves the chemical bonds which hold them together. In a gas, they zoom around freely, crashing into each other once in a while, but always bouncing back rather than bonding. The translational energy associated with this continual motion is what we measure and perceive as temperature. When we put energy into an object to raise its temperature we increase these internal motions. When we extract thermal energy from an object we decrease these internal motions.

Atoms are invisible, and their motion is not generally apparent. Because of this, it took much of the 19th century for scientists to fully understand how the internal energy we associate with temperature was related to the macroscopic forms of kinetic and potential energy. We begin today by talking about collisions. Studying collisions will help us to understand the interactions among atoms, especially in gases. It will help us to see how the atoms in a gas share energy by trading it back and forth in collisions. To understand a gas, it's useful to consider collisions in general.

The physics of collisions has other irresistibly interesting applications in science and technology, and we will also digress a bit in this chapter to explore them. This digression will include the motions of planetary probes and rockets, and an example of how momentum conservation is used in particle physics.

11.2 Collisions between macroscopic objects

What is a collision? A collision is an event (a period of time) during which several objects, isolated from others, interact relatively strongly. The interaction between these objects during the collision should be much stronger than the interaction between either object and anything else. During the collision, only the interaction between the two objects is important for the motion of either.

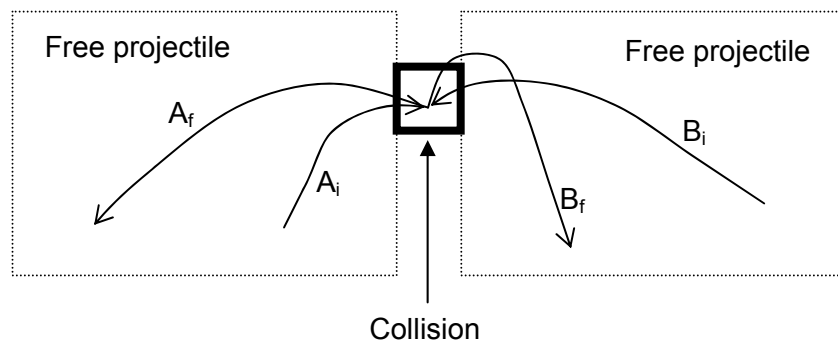
Why study collisions? Mostly because they happen a lot; interactions are often ‘contact’ interactions which occur only for short times. While they are occurring, they are usually vastly more important than other forces in the problem, making the collision approximation a good one. There are many examples of collisions we might use.

When a ball falls to the floor and bounces back up, there is a short time during which the ball is in contact with the floor. During this short time, the forces between the ball and the floor are very large, often much larger than the ball’s weight. This short period is a collision between the ball and the floor. When a softball player hits a pitch, this too is a collision. During the short period when the bat strikes the ball, the forces between the two are very large, much larger than any other forces which act. This short period is a collision too.

Not all collisions are so brief. When a long period comet falls from the outer solar system toward the Sun, it may go through a period of months during which the gravitational force of the Sun on it is much larger than any other force acting on it. This too is a collision.

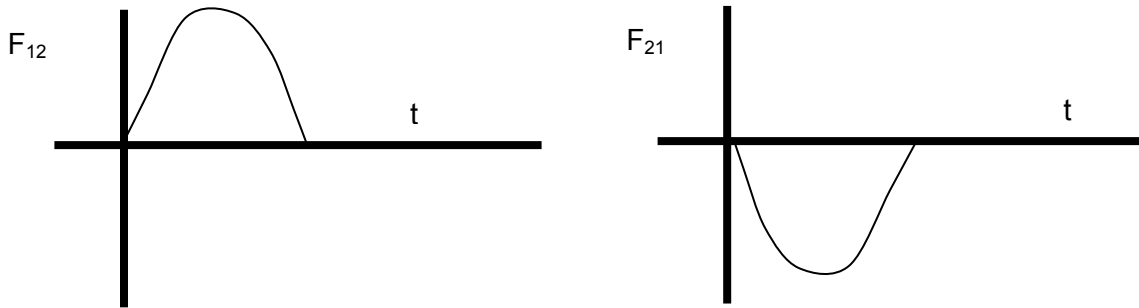
Atoms or molecules in a gas actually spend the vast majority of their time flying quite freely. Occasionally, they get close enough to one another to interact strongly. These interactions take a very short time, and are pretty classic collisions. Once we work out the details of collisions in more tangible systems, we will return to this most important application; understanding the behavior of atoms.

Consider the case of two projectiles colliding in midair. Before the collision their motions are independent and affected only by gravity. During the brief period of the collision the interaction between them is very strong. During that time, the force due to gravity can be briefly ignored; the forces associated with the collision are much larger than the force of gravity. After the collision they again move as free projectiles.



By definition, the only important forces during the collision are those between the two objects, so the only thing which alters their momentum is the force which acts between them. By definition all other forces are unimportant there.

While the two objects collide we know that $\vec{F}_{12} = -\vec{F}_{21}$, just from Newton's third law. This idea is expressed schematically in the figure below.



How do these forces alter the momenta of the two objects? The impulse-momentum theorem tells us:

$$\vec{I}_1 = \int \vec{F}_{21} dt \quad \text{and} \quad \vec{I}_2 = \int \vec{F}_{12} dt$$

Since the forces are equal and opposite, we know the two impulses, and hence the two momentum changes, are equal and opposite. So:

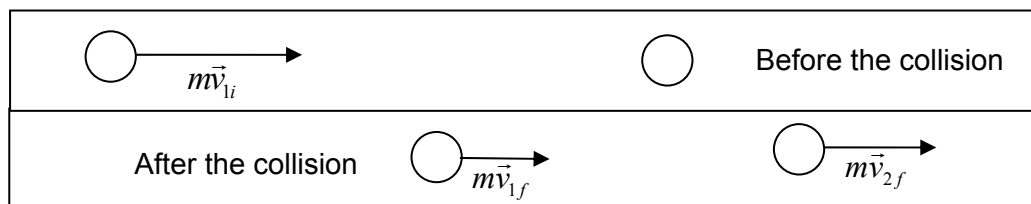
$$\vec{I}_1 + \vec{I}_2 = \Delta\vec{p}_1 + \Delta\vec{p}_2 = 0$$

Whatever momentum change particle one experiences, particle two experiences and exactly equal and opposite momentum change. When you add up the two momentum changes, they perfectly cancel, and the total momentum doesn't change at all. In such a collision, momentum may be transferred from one object to the other, but the total momentum remains unchanged by the collision.

Momentum is *always* conserved in collisions. It doesn't matter at all what kinds of forces act. This is an important point to remember.

A simple example: collision of two identical balls

We begin with a simple example, the collision of two identical balls, each with a mass m . In this example, one of the balls starts out moving with velocity \vec{v}_{1i} , while the other starts at rest. After the collision, the two balls move with velocities \vec{v}_{1f} and \vec{v}_{2f} .



What do we know about this collision? We have seen above that momentum is always conserved in collisions, and in this case we have specified that motion occurs only along the x-axis. Using these facts, we can write a simple momentum conservation equation which says:

$$mv_{1i} = mv_{1f} + mv_{2f} \quad \text{OR} \quad v_{1i} = v_{1f} + v_{2f}$$

So what is v_{1f} ? We have here a single equation describing the motion, but have two unknowns: v_{1f} and v_{2f} . It's impossible to say how the two balls move after the collision unless we have additional information. Even in this extremely simple case we can't determine the outcome of the collision using **only** momentum conservation. We need to know something more.

One kind of additional information we might add to describe a collision is a statement about what happens to energy during the collision. What could we say about this collision if we knew that the total kinetic energy of the system was conserved? In this case, we could also write an additional constraint on the motion:

$$\frac{1}{2}mv_{1i}^2 = \frac{1}{2}mv_{1f}^2 + \frac{1}{2}mv_{2f}^2 \quad \text{OR} \quad v_{1i}^2 = v_{1f}^2 + v_{2f}^2$$

With this additional information, we now have two statements; that the initial velocity equals the sum of the final velocities and that the *square* of the initial velocity is equal to the sum of the squares of the final velocities. Now that we have two equations and two unknowns, we can solve for the two final velocities.

These two equations can only **both** be true if either v_{1f} or v_{2f} is equal to zero. So unless ball one moves through ball two, we must have $v_{2f} > v_{1f}$, and the only solution is:

$$v_{2f} = v_{1i} \quad \text{and} \quad v_{1f} = 0$$

This is the solution for the special case where both momentum and kinetic energy are conserved in a collision which happens along a single straight line. The incoming ball stops and the other takes up its full velocity in its place.

Elastic and totally inelastic collisions

This example illustrates the problem; to understand any collision, we have to know something more about the motion than that momentum is conserved. In simple two body collisions like this there are two extreme cases to consider:

1. “Elastic” collisions: in these the forces which act between the two objects are purely conservative (like springs), so the kinetic energy before and after the collision remains the same. Note that *during* the collision, energy which entered as kinetic energy is often briefly stored as potential energy. The very name ‘elastic collision’ should evoke this thought. When two billiard balls collide on a pool table, the kinetic energy they bring into the collision is (during the collision) briefly stored as elastic deformation of the balls. But they very quickly reconvert that elastic potential energy back into kinetic energy as they spring apart.
2. “Totally inelastic” collisions: in these the forces which act are purely nonconservative (like friction). For these collisions the change in kinetic energy from before to after the collision is as large as possible.

Of course, most collisions lie between these extremes, and hence might be called partially inelastic. In these collisions some of the initial kinetic energy is lost in the collision, but less than the maximum which could be lost.

How can we maximize the kinetic energy lost in a linear collision like this (with two identical objects)? From momentum conservation we know that:

$$v_{1i} = v_{1f} + v_{2f} \quad \text{or} \quad v_{1f} = v_{1i} - v_{2f}$$

and

$$\begin{aligned}\Delta KE &= \frac{1}{2}mv_{1i}^2 - \frac{1}{2}mv_{1f}^2 - \frac{1}{2}mv_{2f}^2 \\ \frac{2\Delta KE}{m} &= v_{1i}^2 - v_{1f}^2 - v_{2f}^2 = v_{1i}^2 - (v_{1i} - v_{2f})^2 - v_{2f}^2 \\ \frac{2\Delta KE}{m} &= v_{1i}^2 - (v_{1i}^2 - 2v_{1i}v_{2f} + v_{2f}^2) - v_{2f}^2 \\ \Delta KE &= mv_{1i}v_{2f} - mv_{2f}^2\end{aligned}$$

We want to maximize the change in kinetic energy by picking the correct v_{2f} . What value for v_{2f} would maximize the change in kinetic energy? We can find this by taking the derivative of the change in kinetic energy with respect to v_{2f} and setting the result equal to zero:

$$\begin{aligned}\frac{d\Delta KE}{dv_{2f}} &= mv_{1i} - 2mv_{2f} = 0 \\ v_{2f} &= \frac{v_{1i}}{2}\end{aligned}$$

To achieve the maximum loss in kinetic energy, we should have ball two move off with a velocity just one half of the velocity ball one arrived with. OK, that's tells us what ball two is doing. What does ball one do in this case? Going back to the momentum conservation equation we started with:

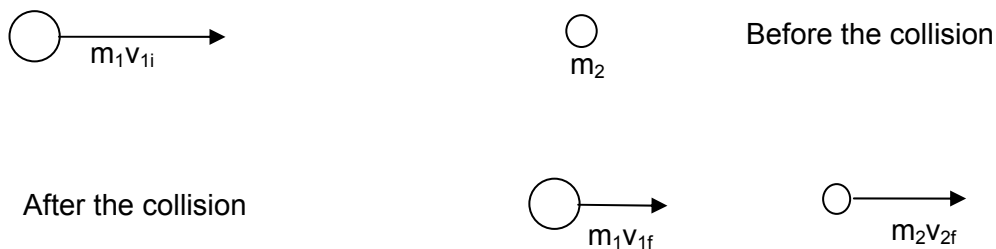
$$v_{1f} = v_{1i} - v_{2f} = \frac{v_{1i}}{2}$$

So both balls move off at the same final velocity. They stick together!

This result, derived here for the simple collision of identical objects in one dimension, is generally true. The way to lose the maximum amount of kinetic energy in a collision is to have the two objects stick together. In totally inelastic collisions, collisions which lose as much kinetic energy as possible while still conserving momentum, the objects always stick together. Remember, they generally don't lose *all* the kinetic energy and come to a stop; most of the time this would violate the conservation of momentum. But when the colliding objects stick together, they do lose the maximum amount of kinetic energy which they can.

Collisions between two balls with different masses

Now consider the same kind of collision in just one dimension, but now let's analyze what happens when the balls which collide have different masses:



Now we can write the momentum conservation equation as:

$$m_1 v_{1i} = m_1 v_{1f} + m_2 v_{2f} \quad \text{or} \quad m_2 v_{2f} = m_1 (v_{1i} - v_{1f})$$

Just as with the case of identical balls colliding, momentum conservation alone does not provide enough information to solve for both v_{1f} and v_{2f} . As before, we can solve for the final motion if we add a statement about what happens to the energy. If, for example, we specify that this is an elastic collision, we can write:

$$\frac{1}{2} m_1 v_{1i}^2 = \frac{1}{2} m_1 v_{1f}^2 + \frac{1}{2} m_2 v_{2f}^2 \quad \text{or} \quad m_1 (v_{1i}^2 - v_{1f}^2) = m_2 v_{2f}^2$$

factoring the left hand side we have:

$$m_1(v_{1i} + v_{1f})(v_{1i} - v_{1f}) = m_2v_{2f}^2$$

Recalling the momentum conservation condition we derived above, we can replace $m_1(v_{1i} - v_{1f})$ with m_2v_{2f} and write:

$$m_2v_{2f}(v_{1i} + v_{1f}) = m_2v_{2f}^2$$

or:

$$v_{2f} = v_{1i} + v_{1f}$$

putting this new condition on the velocities back into the original momentum constraint we find:

$$m_2(v_{1i} + v_{1f}) = m_1(v_{1i} - v_{1f})$$

We can use this to solve for v_{1f} in terms of v_{1i} and the masses:

$$v_{1f} = \left(\frac{m_1 - m_2}{m_1 + m_2} \right) v_{1i}$$

and then use this to find v_{2f} :

$$v_{2f} = \left(\frac{2m_1}{m_1 + m_2} \right) v_{1i}$$

Think back through what we just did. For a one-dimensional collision between two unequal mass balls, we found that momentum conservation alone was not enough to predict what would happen; no surprise there. To solve the problem, we needed to add some information about what happened to energy in this collision. If we specify an elastic collision which conserves kinetic energy, we can solve for the final velocities of both balls.

To get an appreciation of what these equations are telling us, it is useful to consider some limiting cases in which we vary the relative mass of the two.

1. What if the incoming ball is much more massive than the ball which starts at rest? In this case, we would write:

$$m_1 \gg m_2 \quad v_{1f} \approx v_{1i} \quad v_{2f} \approx 2v_{1i}$$

2. If the masses of the two balls are actually equal, we should recover the results we got on in the previous section, which we do:

$$m_1 = m_2 \quad v_{1f} = 0 \quad v_{2f} = v_{1i}$$

3. Finally, what if the mass of the second ball is much greater than the mass of the incoming ball? In this case we would find:

$$m_1 \ll m_2 \quad v_{1f} \approx -v_{1i} \quad v_{2f} \approx 0$$

The third case describes what happens when a ball bounces off something much more massive than it is (like a wall). Since the mass of the incoming ball is so small, its impact is unable to significantly change the motion of the object it strikes. The incoming ball just bounces straight back, leaving the object it strikes sitting there essentially unchanged.

The first case, where m_1 is much larger than m_2 , is perhaps more surprising. In this case the incoming ball continues on essentially as if nothing had happened, while the tiny little second ball, which started at rest, jumps forward, moving with essentially twice the velocity of incoming ball. It actually jumps ahead of the incoming ball.

There are cases where you've probably seen this happen, in a variety of sports. Picture in your mind what happens when you serve a ping-pong ball with a paddle. The paddle is swinging forward at some speed, which continues essentially unchanged after the collision. The ball begins essentially at rest, then "jumps off" the paddle, moving ahead of it with a speed which (you now know) is about twice the speed of the paddle. The same sort of thing happens in a variety of other sports; in baseball when the bat strikes the ball, in football (either kind) when the kicker's foot strikes the ball, and in golf, when the massive club head impacts the golf ball.

Understanding the collision of a massive object with a light one is much more important to us than these sports examples alone might suggest. It helps us to understand what happens when a piston compresses a gas of atoms. In the gas, atoms are always rattling around, bouncing off the walls. Normally, the walls are at rest, and atoms bounce off them with exactly the velocities they arrive with. But if the wall is actually a piston, and it is moving toward the atoms, atoms bounce back (like the ping-pong ball) with a larger velocity than they came in with. As we will discuss in more detail soon, the increase in velocity they acquire in this way corresponds to a change in the temperature of the gas. This is why the temperature of a gas increases when you compress it.

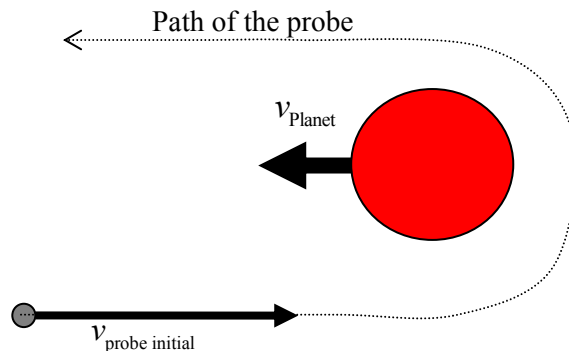
Collisions involving non-contact forces like gravity

You don't have to have contact to have a collision. You only have to have an interaction, and that interaction can take place through a non-contact force like gravity. Imagine what happens when an interplanetary spacecraft like NASA's Cassini Equinox mission, currently in orbit around Saturn, encounters a planet while traveling through the Solar system.

Initially it is flying freely in orbit around the Sun. Then when it nears a planet, the planet's gravity rather quickly, though briefly, becomes much stronger than the gravitational pull of the Sun. During this time, the Cassini mission is undergoing a 'collision' with the planet. In fact, tenuous collisions of this kind are

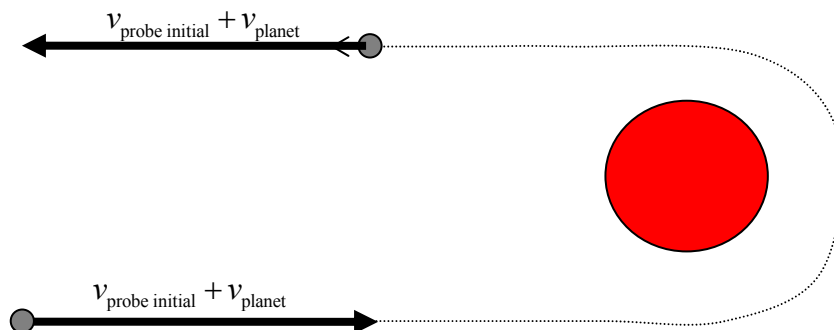
absolutely essential to interplanetary probes. We will work out the basics of these unusual collisions first, then come back to see why the maneuvers are so important.

Consider the collision of an interplanetary probe with a planet. Gravity is all that acts, and since it is a conservative force, this is sure to be an elastic collision. Let us imagine that things have been carefully arranged so that the probe enters from the left, swings around the planet, which is moving to the left, and then departs off to the left again. This simple case is a linear collision of the kind we discussed above, and this will allow us to better see what might happen in such a case. Here is a picture of the probe's path:



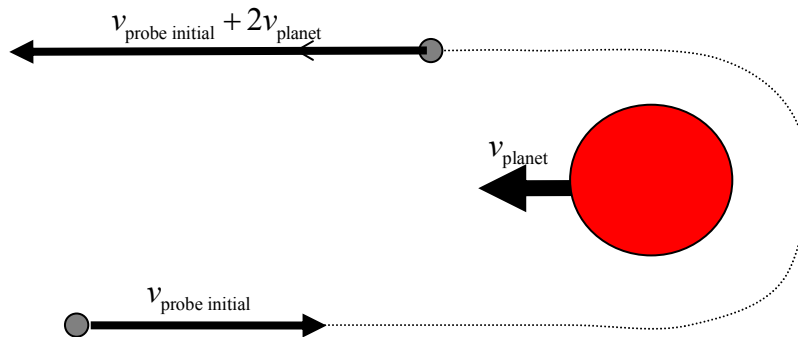
What is $v_{\text{probe final}}$, after the collision when the probe is once again far away from the planet? It doesn't appear at first that this example is very similar to the problems we did above. In those cases, one of the two objects began at rest, while in this case both the probe and the planet are moving before the collision. This question provides us with a good opportunity to use a clever thought experiment to examine an otherwise very complex problem in a simple way.

Imagine how you would view this collision if you were moving through space along with the planet, perhaps because you were sitting on it. From this vantage point you would see the planet sitting still, while the probe came toward you with a larger velocity: $v_{\text{probe initial}} + v_{\text{planet}}$. After the collision, the planet's motion would scarcely be altered by the gravity of the puny probe, and the probe would swing around and head back out the way it came. Since the collision would be elastic, conserving kinetic energy, the probe would leave with just the same velocity it came in with: $v_{\text{probe initial}} + v_{\text{planet}}$. So here's what the collision looks like when viewed from the reference frame of the moving planet:



How do we know that the satellite will go back out with a velocity just as large as it came in with? This collision, viewed from the planet, is just the same as the one we analyzed above. Seen in this way, we have a very low mass incoming object striking a very massive second object which is initially stationary. As we saw before, in this case, the incoming object just bounces straight back out.

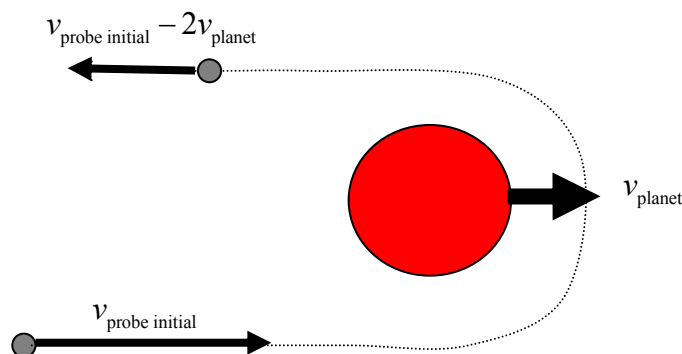
The NASA engineer, however, doesn't care about the velocity of the probe relative to the planet. She wants to know the velocity of the probe relative to some global solar system standard of rest. We can find this from the above diagram by imagining what it would look like if we weren't moving along with the planet. In that case, we would see:



Now you can see how the probe gains speed in this collision. You might also see a close similarity to the ping-pong ball on paddle collision we worked out before. In this case, the ball left the collision gaining twice the velocity of the paddle. That's all that is happening here. This kind of collision with a heavy particle can be used to accelerate a light one in all kinds of cases. The same thing happens when electrons collide with atomic nuclei in a process known as 'inverse Compton scattering'. This happens in the very hot plasmas which surround exotic astrophysical objects like black holes and neutron stars. Collisions like this can accelerate electrons to enormously high energies in such conditions.

This trick, recognizing that once you solve the collision in which one object begins at rest you have solved all collisions, was first recognized by Christian Huygens. It was an important part of his initial efforts to understand collisions and introduce the idea of kinetic energy.

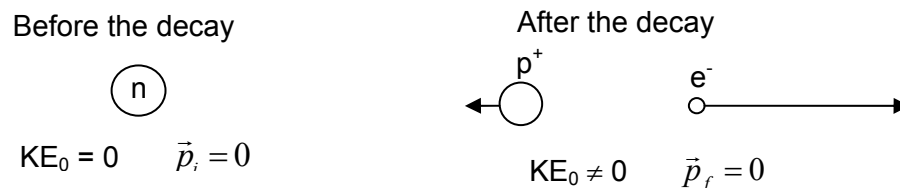
It is interesting to also think about the reverse process. When the probe approaches a moving planet from behind, it actually loses velocity in the collision.



What happens here is a kind of braking. These two techniques are called gravity ‘assists’. NASA and the European Space Agency (ESA) use these gravity assists to accelerate interplanetary probes up to the speeds they need to travel through the solar system. Without gravity assists of this kind it would be practically impossible for us to get satellites to the outer planets like Jupiter and Saturn; we would need to use rocket fuel alone to accelerate our probes. They also use the reverse process (the gravity braking) to slow satellites down when they reach their destinations, though more often this is done using friction with the atmosphere of the planet.

An inverse collision: 2 body decay and the neutrino

As a final example, consider the decay of a neutron. It turns out that if you leave a neutron sitting around by itself, it will eventually fall apart; decaying spontaneously into less massive pieces. In 1928 this was thought to involve a neutron turning into just a proton and an electron. A decay is rather like an inverse collision. Instead of several objects coming together and sticking, one object comes apart into pieces. Here is a picture of how this decay was envisioned in 1928.



It was known at the time that the energy released in this decay came from “mass energy”, the kind of energy described by Einstein’s most famous equation $E = mc^2$. Since the mass of the neutron is a little bigger than the sum of the mass of the proton and the mass of the electron, energy is available in the neutron to release into the proton and the electron after the decay. The amount of available energy is:

$$E_{\text{decay}} = m_{\text{neutron}} c^2 - (m_{\text{proton}} + m_{\text{electron}}) c^2$$

We know that momentum is conserved. Scientists at the time thought that since this decay should have involved fundamental interactions, energy should be conserved too, so that the mass energy which was released in the collision should all show up as kinetic energy in the proton and electron. If this were the case, we could use momentum conservation to assert that the proton and electron must come out back-to-back and write:

$$0 = m_{\text{electron}} v_{\text{electron}} - m_{\text{proton}} v_{\text{proton}}$$

$$v_{\text{proton}} = \left(\frac{m_{\text{electron}}}{m_{\text{proton}}} \right) v_{\text{electron}}$$

Momentum conservation alone provides a specific prediction for the relation between the electron and proton velocities. If the kinetic energy which emerges all comes from the mass energy of the decay, then we would also know:

$$E_{\text{decay}} = \frac{1}{2} m_{\text{proton}} v_{\text{proton}}^2 + \frac{1}{2} m_{\text{electron}} v_{\text{electron}}^2$$

Putting these two equations together, we can solve for the energy of the electron and proton and find:

$$E_{\text{decay}} = \frac{1}{2} m_{\text{proton}} v_{\text{proton}}^2 + \frac{1}{2} m_{\text{electron}} \left(\frac{m_{\text{proton}}}{m_{\text{electron}}} \right)^2 v_{\text{proton}}^2 = \frac{1}{2} m_{\text{proton}} v_{\text{proton}}^2 \left(1 + \frac{m_{\text{proton}}}{m_{\text{electron}}} \right)$$

$$E_{\text{proton}} = \frac{E_{\text{decay}}}{\left(1 + \frac{m_{\text{proton}}}{m_{\text{electron}}} \right)}$$

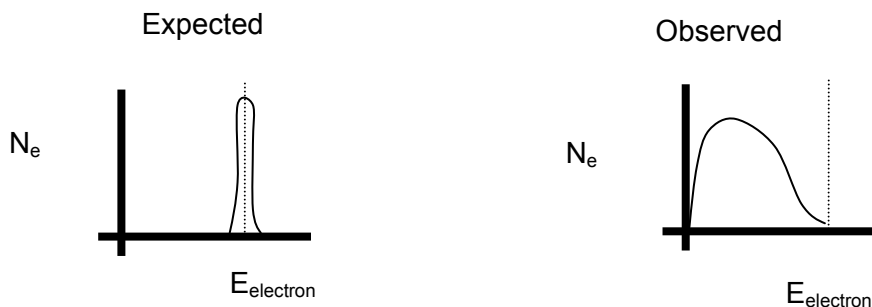
Similarly for the electron:

$$E_{\text{decay}} = \frac{1}{2} m_{\text{proton}} \left(\frac{m_{\text{electron}}}{m_{\text{proton}}} \right)^2 v_{\text{electron}}^2 + \frac{1}{2} m_{\text{electron}} v_{\text{electron}}^2 = \frac{1}{2} m_{\text{electron}} v_{\text{electron}}^2 \left(1 + \frac{m_{\text{electron}}}{m_{\text{proton}}} \right)$$

$$E_{\text{electron}} = \frac{E_{\text{decay}}}{\left(1 + \frac{m_{\text{electron}}}{m_{\text{proton}}} \right)}$$

These relations must be true if momentum is conserved and all the mass energy released in the decay appears as kinetic energy in the proton and electron. It says that if you measure the energy of electrons emerging from such a decay, you should find just one value for their energy. Since the mass of the electron is about 1800 times smaller than the mass of the proton, these equations suggest that almost all of the decay energy should end up in the electron, while very little ends up in the proton.

This was a nice experimental prediction, and was pretty promptly put to the test. The results were a great surprise. Instead of every electron emerging with the predicted energy, they all came out with different energies; from nearly zero all the way up to a maximum which coincided with the energy predicted from the above calculation. The situation is illustrated in the cartoon histograms shown below.



This was a big problem for physics. How could this be explained? At the time, there seemed to be three choices:

1. Momentum is not conserved (a terrible possibility)
2. Energy is not conserved (even more terrible)
3. There is a third (undetected) body in the collisions (surprising, but nicer...)

Wolfgang Pauli, one of the leaders among early quantum theorists, originally suggested the presence of this third, electrically neutral object in such decays. With three bodies in the decay, E_{electron} can be anything from 0 to E_{decay} . This ‘missing’ particle would then take up whatever energy and momentum was left over.

Enrico Fermi worked out the details of the predictions for this three body decay, and called the new particle a ‘neutrino’ because it was a little neutral one and he was Italian. The neutrino was not observed independently for 27 years after Fermi’s prediction of it. Why so long? Because neutrinos hardly ever interact with anything. Right now there are about 100 billion neutrinos per square centimeter passing through your body each second. They flow out from the Sun at a truly extraordinary rate. Essentially none interact; they go straight through you as easily (much more easily really) as light passes through a window.

How then do we know neutrinos are real? Well, they sometimes *do* interact. By building large enough experiments, and watching carefully enough, we can observe them directly. But their existence was revealed by analyzing the decay of the neutron and insisting that momentum and energy should be conserved.

Another inverse collision: rocket propulsion

The decay of the neutron opens the door to an examination of rocket propulsion. Notice what happened in this decay. The neutron broke apart into a proton and an electron, releasing some energy stored inside it to send the electron flying off in one direction and the proton flying off in the other.

You might use the same propulsion approach by standing on a good skateboard, then throwing a ball backward. When you exert an impulse to send the ball flying off in one direction, it exerts an equal and opposite impulse to send you flying off in the other direction. We know that momentum will be conserved in this interaction, so we could write:

$$m_{\text{you and skateboard}} v_{\text{you and skateboard}} = m_{\text{ball}} v_{\text{ball}}$$

Imagine that the speed with which you throw the ball relative to you is v_{throw} . You might find this by standing firm on solid ground and throwing the ball. Since this is the total velocity you can create between you and the ball, we can also write:

$$v_{\text{throw}} = v_{\text{ball}} + v_{\text{you and skateboard}}$$

Solving these two equations allows us to write your speed and the ball’s speed relative to the ground as:

$$v_{\text{you and skateboard}} = \left(\frac{m_{\text{ball}}}{m_{\text{you and skateboard}} + m_{\text{ball}}} \right) v_{\text{throw}}$$

$$v_{\text{ball}} = \left(\frac{m_{\text{you and skateboard}}}{m_{\text{you and skateboard}} + m_{\text{ball}}} \right) v_{\text{throw}}$$

Notice what this implies. If you and the skateboard weigh much more than the ball, then $v_{\text{you and skateboard}} \ll v_{\text{throw}}$ and $v_{\text{ball}} \cong v_{\text{throw}}$. If the ball you throw has a mass equal to yours, both you and the ball will move off with half the speed of the throw (relative to the ground). So if you want to get yourself moving by throwing something away from you, you should be sure to throw something as massive as possible, and to throw it as hard as possible (with the largest possible v_{throw}). This approach is exactly how rockets work, though they tend to throw fuel out the back in a continuous stream, rather than one object at a time.

The challenge of reaching large velocities with rockets is serious, for the following reason: when a rocket starts out, blasting its first little bit of fuel out the back, it gets an initial impulse. But that impulse is applied to a rocket which contains not only the payload you're interested in speeding up, but also all the rest of the fuel required for the trip. Much of the rocket's initial impulse goes into speeding up fuel, rather than the payload.

To give a sense of how serious a problem this is, note that the US space shuttle requires an initial mass of fuel which is about 20 times the total mass of the shuttle and its cargo: 2,000,000 kilograms at lift-off as opposed to 100,000 kilograms at landing. And that's just to get to near-Earth orbit, just a few hundred kilometers above the surface. To get to the Moon the Apollo 11 mission used a Saturn V launch vehicle which weighed fully 65 times as much as the spacecraft it was intended to launch.

These figures should make it clear why gravity assists, in which a spacecraft can gain velocity using no fuel at all, are so important for space exploration. Without them, we would not yet have explored most of the solar system.

11.3 Collisions in 3D

So far we've only worried about collisions in one dimension. What happens if we try to extend this analysis to more typical 3D collisions? What do we need to know to completely describe a collision between two objects in three dimensions? The list of things we need to know would include:

1. The masses of the two objects (2 parameters)
2. The initial velocities of each object (3 components for each object, 6 total parameters)
3. The final velocities of each object (another 6 parameters)

So there are 14 parameters we need to know to fully describe such a collision. We might imagine being given the initial conditions: the masses and the initial velocities of both objects. This would be a total of 8 parameters, and we would still need to determine the other six. How could we do this?

We begin by assembling what we know about collisions. Momentum is conserved in all collisions, so we can always write:

$$m_1 \vec{v}_{1i} + m_2 \vec{v}_{2i} = m_1 \vec{v}_{1f} + m_2 \vec{v}_{2f}$$

Because this is a vector equation, it must hold true for each of the three components of the momentum. So the momentum conservation equation yields three constraints on the velocities. But we need to determine 6 unknowns. Just as in the linear cases we considered above, momentum conservation alone is not enough to allow us to make a prediction. We need to know something more.

We might know, for instance, that the collision is elastic, so that the kinetic energy before and after the collision are the same. This would provide one new constraint on the velocities. Clearly this would not be enough for us to solve the full 3D problem.

Let's look at an example, a collision between two objects with masses m_1 and m_2 . To keep things simple, imagine that m_1 begins moving along the x-axis, and m_2 begins at rest, and that the final motion is restricted to the xy-plane.



Imagine we are given here: m_1 , m_2 , and v_{1ix} and told that

$$v_{1iy} = v_{1iz} = v_{2ix} = v_{2iy} = v_{2iz} = v_{1fx} = v_{2fx} = 0$$

So we're given the ten things. Momentum conservation in the x and y directions gives us two equations:

$$m_1 v_{1ix} = m_2 v_{2f} \cos(\theta_2) + m_1 v_{1f} \cos(\theta_1)$$

$$0 = m_1 v_{1f} \sin(\theta_1) + m_2 v_{2f} \sin(\theta_2)$$

And if we also know that the collision is elastic, so that kinetic energy is conserved, then we have another equation:

$$\frac{1}{2} m_1 v_{1ix}^2 = \frac{1}{2} m_1 v_{1f}^2 + \frac{1}{2} m_2 v_{2f}^2$$

But this is 13 things either given or constrained by equations, when the full description of the collision includes 14 unknowns. Even in this simple case, we don't have enough information to fully predict what will happen when two objects collide.

What **can** we say about this? This general problem can't be solved. But we can still make predictions. If we specify a part of the outcome, we can predict the rest. For example, we might specify that m_1 emerges at an angle $\theta_1 = 30^\circ$. If we did this, we would then have enough information to completely specify all the remaining unknowns, including the final velocities of both objects and the angle θ_2 which describes the direction of the second mass. Very often, then, collisions like this are solved by understanding the distribution of results which will occur. Whenever we see mass one emerge with angle θ_1 , we know for sure what the final velocity of mass two will be.

So basically I can't tell you what will happen in a collision without more information than just momentum and energy conservation. But I can tell you that if you see particle 1 come out at an angle θ_1 , you will completely determine the motion of particle 2.

11.4 Collisions of atoms and molecules in gases

Our primary reason to examine collisions in so much detail is to begin exploring the internal energy present in the motion of the atoms which make up a material. A good, clean example of this is the situation experienced by a typical gas (like air) at room temperature. In such a gas, atoms and molecules spend most of their time flying freely. You shouldn't get the impression that they don't interact often, for they do, typically traveling less than a millionth of a meter between collisions and running into other atoms several billion times a second. Nevertheless, the interactions between atoms are very brief indeed, so that much more time is spent flying between collisions than engaged in them.

What happens in such a gas? There are a truly huge number of collisions each second, and in each, momentum and energy is exchanged between the atoms. The very large number of collisions insures that atoms with more energy and momentum than average will quickly share this out among their neighbors. Atoms which have less energy and momentum than average will quickly gain some from interactions with their more energetic neighbors. The net result, which we will explore in more detail in Chapter 13, is a very rapid approach to a stable mix of particle energy and momentum. Collisions allow the atoms to share momentum and energy until they quickly reach a stable distribution; some have more, some less. Each changes constantly, but the overall distribution of fast and slow quickly settles down and remains the same.

Are collisions between atoms and molecules elastic, or can they store energy internally?

When two atoms or molecules collide, are the collisions elastic, so that kinetic energy is conserved in the collision? Or are they partially inelastic (so that some of the incoming energy is lost) or totally inelastic (so that the two stick together after the collision). All of these outcomes are possible, in the right circumstances. But for materials which are gases at room temperature, it is quite common for all of the collisions among atoms and molecules to be elastic. The reason for this lies in the nature of the quantum mechanics which governs the behavior of atoms and molecules.

Quantum mechanics describes the behavior of very small objects, like atoms and molecules. A common feature of such systems is that they can only possess only a finite set of states; each possessing specific and often widely separated energies. The lowest energy state of an atom is called its ground state, the next highest its first excited state, and so on. Imagine that two atoms, each in their ground states, collide. If the total energy in the collision is *less* than the energy difference between the ground and first excited state, it is impossible for the atoms to absorb the available energy and remove it from their motions. Since the energy has nowhere else to go, it must remain in the motion of the atoms, and the collision will be elastic.

If the collision energies are larger than the gaps between available quantum states then indeed the collisions may be at least partially inelastic, with some of the energy initially present in motion being absorbed into internal energy states within the atoms or molecules. As we will see in the next chapter, if you want to get a gas with such internal energy states to begin moving with a larger average kinetic energy, you will have to add more energy than you might expect, since some of what you put in will leak away into these other internal energy states.

What if the collisions between the atoms and molecules in a gas are totally inelastic? In this case, we know they will stick together, losing the maximum amount of kinetic energy. When the atoms or molecules in a gas stick together, they may stop bonding in small groups - as might happen if you could mix monatomic N atoms together. They would quickly merge into a gas of N₂ molecules. If many atoms can stick together, your gas may quickly condense, collapsing from a gas into a liquid or solid.

From this you can see why the substances which are gases at room temperature must not undergo these kinds of totally inelastic collisions. If they did the substance would not be a gas in the first place.

A Quick Summary of Some Important Relations

Essentials of collisions:

A collision is an interaction between two objects which lasts for a limited time. When the only unbalanced forces acting on the objects are the ones they exert on each other, momentum will *always* be conserved in a collision.

$$\begin{aligned}\vec{p}_{1i} + \vec{p}_{2i} &= \vec{p}_{1f} + \vec{p}_{2f} \\ \Delta\vec{p}_1 + \Delta\vec{p}_2 &= 0 \\ \Delta\vec{p}_1 &= -\Delta\vec{p}_2\end{aligned}$$

These equations emphasize that in a collision, momentum is transferred from one object to another: it is never lost, only transferred.

Kinetic energy in collisions:

Kinetic energy may be conserved in a collision, in which case we call it ‘elastic’, or some kinetic energy may be lost. If the collision suffers the largest possible kinetic energy loss, we call it ‘totally inelastic’. This happens when objects stick together, losing all relative motion. Even when kinetic energy is conserved, it is usually traded between the two colliding objects – transferred from one to the other like the momentum.

One dimensional collisions:

In 1D collisions momentum conservation and some information about kinetic energy loss is enough to fully define the problem. You should be able to predict the final motions of two objects in cases like this.

Two and three dimensional collisions:

In these cases, momentum conservation and a single statement about energy is not enough to fully define the outcome; some additional condition must be specified.

Collisions of atoms and molecules in gases:

The most important application of this analysis of collisions is to the motions of atoms in a gas. In this case, incredibly numerous collisions allow the atoms to share out any energy and momentum they possess. This process quickly spreads the energy through the gas, and leads to a stable mix of velocities among the atoms.

12. Mixing it up: oscillations as a mix of kinetic and potential energy

- 1) Describing harmonic motion
 - i. Oscillations around equilibrium
 - ii. Simple harmonic motion
 - iii. Linear restoring forces cause simple harmonic motion
 - iv. Natural frequencies of oscillation
- 2) Examples of simple harmonic motion
 - i. A mass on a linear spring
 - ii. The simple pendulum (for small oscillations)
 - iii. An oscillating rod: hair cells in the ear
- 3) Trading potential and kinetic energy in harmonic motion
 - i. Small oscillations are always simple harmonic motion
 - ii. An example: intermolecular bonds
- 4) Damped harmonic motion
 - i. Weak damping
 - ii. Overdamped motion
 - iii. One specific model for damped oscillators
 - iv. Natural frequencies of oscillation
- 5) Driven oscillations and resonance
 - i. Familiar examples of resonance

Physics for the Life Sciences: Chapter 12

12.1 Describing harmonic motion

In this chapter we're going to spend time describing in some detail the motion of objects which oscillate around an equilibrium position. This kind of oscillation happens whenever an object held in place by some force - like an atom in a solid - is nudged away from its equilibrium position. When this happens, it gets pulled back toward that position of rest, often overshooting it and then oscillating back and forth around it.

Why is this kind of motion important for life?

The most basic reason is that it is intimately connected with temperature and thermal energy. When we say that a solid is hot, and has a temperature T , what we mean is that the atoms in that solid are moving. They aren't going anywhere on average, they're just oscillating around their average rest positions. If the temperature is higher, they're moving more, and in fact the average energy associated with this shaking around is directly proportional to the temperature. If we're going to understand the flow of energy through living systems, which is what makes all of life possible, we need to understand a lot more about this.

There is another important aspect of oscillating systems relevant for life. When oscillating systems are disturbed by outside forces, they exhibit a behavior called 'resonance'. Their response to action from the outside is selective. If you disturb them in just the right way you may get an extremely strong response, while other disturbances have little impact. This phenomenon of resonance is used in many ways by living things, so it's another aspect of oscillations we need to introduce.

Most motions in the real world exhibit neither constant velocity nor constant acceleration. In fact most objects spend their time sitting at equilibrium. Motion of an object around equilibrium will be oscillatory. We'll review why in a minute. To describe this we need to talk a bit about the 'kinematics' of harmonic motion. What kind of language do we have to use to *describe* motion which repeats over and over?

Any motion which repeats itself we will call harmonic (or periodic) motion. Examples include:

- Pendulum or swing
- Ball on spring
- Rotation of Earth
- Orbit of moon
- Motions of and sounds produced by musical instruments
- Motion of atoms in solids
- Water waves

The basic feature of all of these things is that they have some measurable property (position, velocity, height, electric field strength, density, pressure, temperature....) which goes through a pattern of change with time that is repeated. In most cases such oscillatory motion doesn't continue forever; it gradually fades away. The energy associated with oscillations is gradually drained away due to frictional losses. We'll work through some ways of including this damping in oscillations later in this chapter. For the moment, we will imagine the oscillations happen without friction.

To describe any kind of periodic motion we have to specify two things about it:

1. The first thing we need is a measure of how often the oscillations occur. We describe this with the **frequency**; the number of times per second that the motion repeats itself. The units of frequency are inverse seconds. These units are called 'Hertz' for Heinrich Hertz, the great German experimentalist who showed that light was a traveling electromagnetic oscillation. In addition to frequency, we will sometimes speak of the **period** of oscillation; the amount of time each oscillation takes. The period has units of seconds and is just the inverse of the frequency. We will usually use the symbol f for frequency and T for period.
2. The second thing we need is a measure of how large the oscillations are. This is encoded in the **amplitude**. The amplitude takes on the units of whatever is oscillating. If we're talking about the motion of a ball on a spring the amplitude is measured in meters. We will usually use the symbol A for amplitude.

This kind of description is completely general, not referring to any particular kind of oscillatory motion, but just to oscillations in general. Anything which is oscillating will do so with some frequency and some amplitude.

Oscillations around equilibrium

An object is at equilibrium any time the net force acting on it is zero. So of course most of the things you see around you are objects at equilibrium. Equilibrium points are those places where an object feels no force. There are three kinds of equilibrium points. You can understand the nature of different kinds of equilibria by thinking about what happens when you nudge the object a small distance away from its equilibrium point.

- Stable: if you move the object a little away from equilibrium, it is pushed back towards it. Examples of objects at stable equilibrium include a ball resting in the bottom of a bowl, or a swing hanging straight down.
- Unstable: if you move the object a little away from equilibrium, it is pushed farther away from it. Examples of unstable equilibrium include a ball resting on the top of a hill, or a pencil carefully balanced on its eraser.
- Neutral: if you move the object a little away from equilibrium, it isn't pushed either back toward or farther away from it. A good example of neutral equilibrium is a ball resting on a flat plane.

Just how unstable or stable the equilibrium is depends a lot on the details of the system, but the nature of each equilibrium points does not.

Simple Harmonic Motion

One particular kind of harmonic motion is called “Simple Harmonic Motion” which we will sometimes abbreviate as SHM. In this kind of motion the position of an object varies sinusoidally with time. We can describe this motion as:

$$x(t) = A \cos(\omega t)$$

Once we know this, we can find the velocity and acceleration of the particle from their definitions:

$$v(t) = \frac{dx(t)}{dt} = -A\omega \sin(\omega t)$$

$$a(t) = \frac{dv(t)}{dt} = -A\omega^2 \cos(\omega t)$$

The position-time, velocity-time, and acceleration-time graphs for this motion are all shown in the figure on the next page. There's something to notice about this. The equation for the acceleration is almost the same as the equation for the position:

$$a(t) = -A\omega^2 \cos(\omega t) \quad \text{and} \quad x(t) = A \cos(\omega t)$$

so

$$a(t) = -\omega^2 x(t)$$

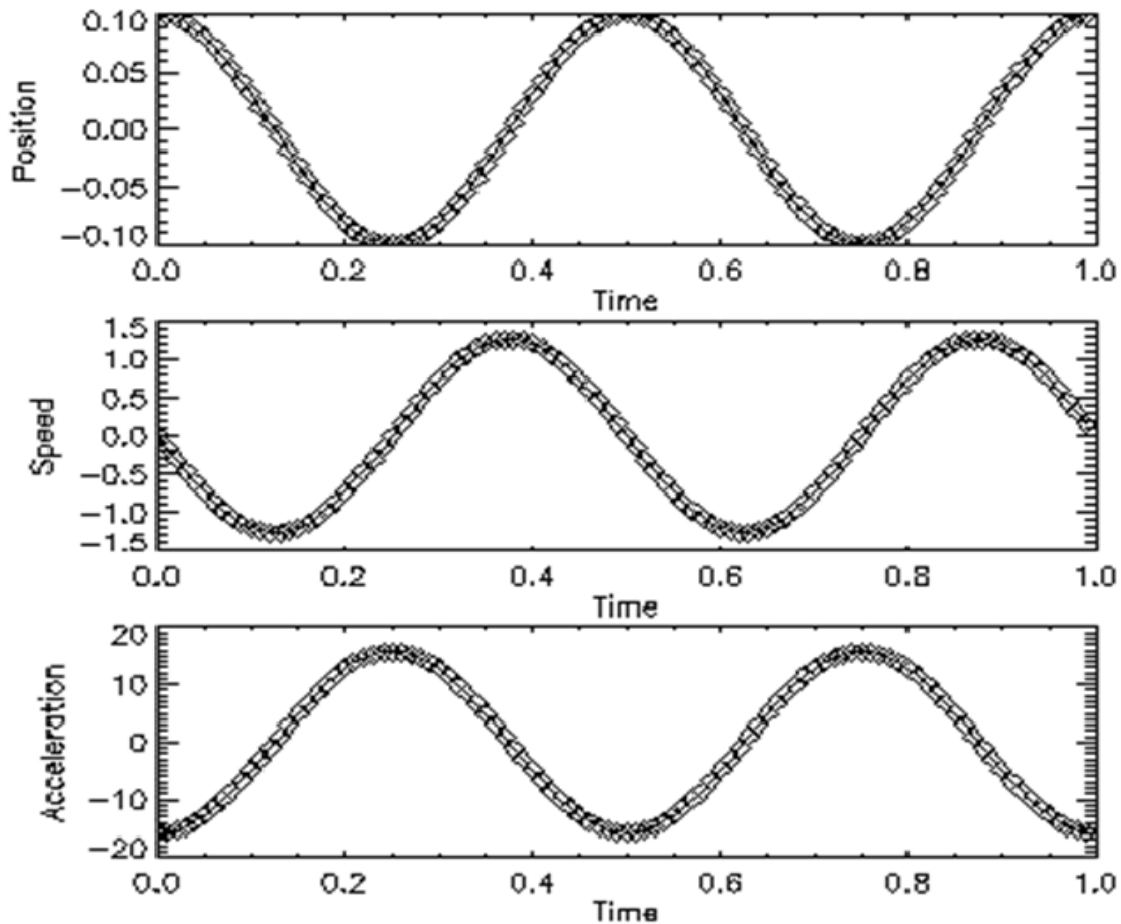
What are the properties of this motion? We said every harmonic motion should have a frequency, period and amplitude. For this motion, the frequency is encoded in this “angular frequency” ω . This is related to the regular frequency according to:

$$f = \frac{\omega}{2\pi}$$

and the period

$$T = \frac{1}{f} = \frac{2\pi}{\omega}$$

How can you see this? When the argument of the cosine function goes from zero to 2π , the value of the cosine goes from 1 through a cycle to -1 and returns to 1. Since the period is the time this takes to happen, $\omega T = 2\pi$, or $T = 2\pi/\omega$. Meanwhile, the amplitude of oscillation is the parameter A in front of the equation.



Here is an example, if the equation of motion is:

$$x(t) = (0.1 \text{ m}) \cos(4\pi t)$$

then the amplitude of motion is 0.1 m. That is, the object oscillates between +.1 m and -.1 m. The angular frequency of oscillation ω is 4π rad/s. From this we can find the frequency of oscillation:

$$f = \frac{\omega}{2\pi} = \frac{4\pi \text{ rad/s}}{2\pi} = 2.0 \text{ Hz}$$

and the period of oscillation is:

$$T = \frac{1}{f} = 0.5 \text{ s}$$

Linear restoring forces cause simple harmonic motion

Remember the Hooke's law force, in which the force returning an object to equilibrium gets larger in direct proportion to the distance from equilibrium. The equation which describes this linear restoring force is:

$$F = -kx$$

We also know from Newton's second law that, if this is the only force that acts, we can write

$$F = ma = -kx$$

Or

$$a = -\left(\frac{k}{m}\right)x$$

Notice that this looks very similar to the equation we wrote down relating position and acceleration for simple harmonic motion:

$$a = -\omega^2 x$$

So, an object moving under the influence of a linear Hooke's law force will move in simple harmonic motion with an angular velocity:

$$\omega^2 = \frac{k}{m} \quad \text{or} \quad \omega = \sqrt{\frac{k}{m}}$$

Notice what this means. An object subject to a Hooke's law restoring force will oscillate rapidly (with high frequency) if the spring constant k is large. If the spring constant is small, the oscillations will be slower. But it is not only the strength of the restoring force which sets the frequency of oscillation. The inertia of the object experiencing the force affects it too. If the mass of the moving object is large, it will slow the frequency of oscillation. If the mass is small, it will oscillate more quickly.

A key point is that the frequency will always be determined by a competition between inertia and restoring force. This general tendency is true even if the restoring force is not linear and the motion not

SHM. Increasing the strength of the restoring force (k) will always speed the oscillations, while increasing the inertia (m) will always reduce them.

Referring to Hooke's law gives the impression that the only time you get simple harmonic motion is with a mass on a spring. This isn't the case at all. In fact any time an object is at a point of stable equilibrium you can get this kind of motion. There are limits of course, and they're basically limits on how big the amplitude of oscillation can be. If you make an atom in a solid oscillate too far from its position of rest, it will break out of the solid and not oscillate at all. But for small enough oscillations, the motion around any position of rest will be simple harmonic motion.

12.2 Examples of harmonically oscillating systems

1: The first example of a system which oscillates harmonically is the mass on spring we have just worked out. For this case the force is the familiar Hooke's law one, and the angular frequency of oscillation is determined by a balance of spring constant and mass.

$$F = -kx \quad \text{and} \quad \omega = \sqrt{\frac{k}{m}}$$

2: A second example of a harmonically oscillating system is the simple pendulum. What makes this particular pendulum 'simple' is that its mass is all at the bottom, a distance L from the point where it is suspended.

In this case we're going to start by considering only small oscillations around equilibrium, where the angle θ never becomes large. In this case we'll be able to use the small angle approximation, which says that for small angles (and remember, this only works for angles measured in radians!):

$$\sin(\theta) \cong \theta$$

How small do the angles have to be for this to be a good approximation? We can check by putting in different values. What we find is that it's a good, better than 5%, approximation to surprisingly large angles, like 30° .

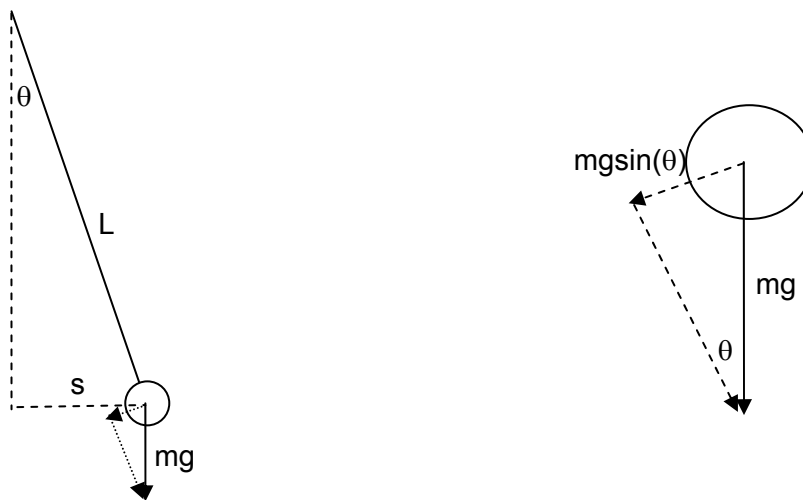
Angle	Sin(θ)	θ	% difference
$\pi/16$ (11.25°)	0.1951	0.1935	0.8%
$\pi/8$ (22.5°)	0.3827	0.3927	2.6%
$\pi/4$ (45°)	0.7071	0.7854	10.0%

With that in hand, let's work out how big the force is which returns the pendulum ball to its rest position. Referring to the drawing below, we can see that the force which pulls it back to the middle is approximately

$$F_{\text{restoring}} = -mg \sin(\theta) \cong -mg\theta \cong -\frac{mg}{L}s$$

Along this line we used both the small angle approximation and the definition of the angle in radians

$$\theta \cong \frac{s}{L}.$$



What about the direction of this force? You can see from the picture that if s is positive (you push it to the right), the force will be negative (it will be pushed back left) and vice versa. This is a restoring force. That's the origin of the minus sign in front of the equation.

This force law looks exactly like Hooke's law, with a 'spring constant' $k_{\text{effective}} = mg/L$. This means the pendulum will show oscillations in the offset parameter s with angular frequency

$$\omega = \sqrt{\frac{k_{\text{effective}}}{m}} = \sqrt{\frac{g}{L}}$$

This result is perhaps a bit surprising, because it suggests that the pendulum will oscillate with the same frequency no matter what mass we put on it. Why? Because both the inertia and the restoring force depend on mass. Any gravitationally restored oscillator will share this property.

Notice some things about what this says; the period of oscillation of a pendulum is *only* dependent on its length. Not on its mass, the material it is made of, or anything; just its length. This is why, throughout your childhood, you could swing on a swingset 'in synch' with your friends, or even your parents,

regardless of how much each of you weighed. All that mattered is that the length of each swing was the same.

3: Another example is the oscillating rod. In this case you “shear” the rod sideways by some small amount Δx , then release it and let it oscillate. Here we can relate the shear stress and strain on the rod in the usual way, through the shear modulus S :

$$\frac{F}{A} = -S \frac{\Delta x}{L}$$

This can be rewritten in a Hooke’s law form as:

$$F = -\left(\frac{SA}{L}\right)\Delta x$$

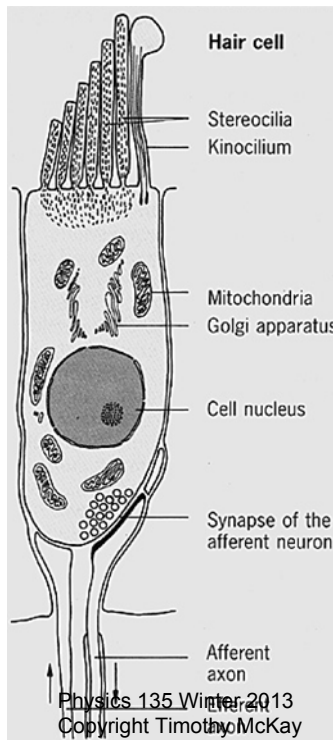
Here you have another effective spring constant $k_{\text{effective}} = \frac{SA}{L}$. This means it will oscillate with:

$$\omega = \sqrt{\frac{k_{\text{effective}}}{m}} = \sqrt{\frac{SA}{Lm}}$$

In fact, we can write this another way, using the fact that $m = \rho V = \rho AL$ to find:

$$\omega = \sqrt{\frac{SA}{Lm}} = \sqrt{\frac{S}{\rho L^2}}$$

So once again, a stiffer rod, with high shear modulus S , will oscillate faster. A longer rod will oscillate more slowly. If the rod is denser, it will also oscillate more slowly. This particular example is used in



your body. Your hearing works through a set of tiny, finely tuned hairs which oscillate as little versions of these rods.

12.3 Energy in harmonic oscillators

Oscillators involve both motion and forces, so to fully understand them we need to think about the flow of energy. When a ball oscillates back and forth on a spring it is repeatedly trading kinetic and elastic potential energy. If there is no friction, this trade can continue indefinitely. When friction is present it will gradually drain away the energy. The motion in this case will be damped, gradually dying away. We will discuss the phenomenology of this damping and develop one quantitative model for it in a bit.

We will also have to talk about how oscillators get started. In particular, we'll talk about the surprisingly important case of an oscillator driven by a force which is itself periodic. This seems like a bizarre and unlikely case, but is in fact remarkably common. It comes about because you so often have one oscillator next to another, as you do with atoms in a solid. If you disturb one atom, it starts to oscillate. In doing this, it bangs repeatedly, and periodically, into its neighbors. This periodic driving found in systems of coupled oscillators is why studying oscillators driven by periodic forces is so important.

We begin by considering a mass oscillating on a spring with no friction. When this mass oscillates it takes energy stored in the elastic stretching of its spring and converts it to kinetic energy, shooting through equilibrium, and then turns that kinetic energy back into elastic potential. This process repeats over and over again.

Given what we know about the motion of such an oscillator, we can quantify its energy quite easily:

$$KE = \frac{1}{2}mv^2 = \frac{1}{2}m \left[A^2 \omega^2 \sin^2(\omega t) \right] = \frac{1}{2}kA^2 \sin^2(\omega t)$$
$$PE = \frac{1}{2}kx^2 = \frac{1}{2}kA^2 \cos^2(\omega t)$$

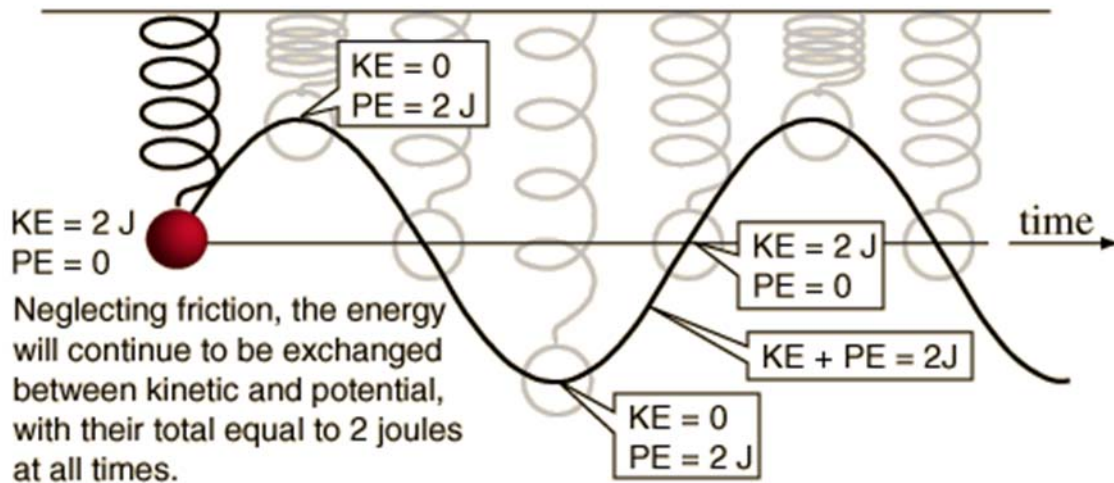
All we did to find these was to plug in the equations described before for $x(t)$ and $v(t)$ and take advantage of the fact the $\omega^2 = k/m$. These equations show just how the energy oscillates from kinetic to potential and back again. Notice that the PE is at a maximum when the position $x(t)$ is at a maximum; when the spring is fully stretched out. The kinetic energy, on the other hand, is at a maximum when the velocity is at a maximum. This happens just when the object passes through equilibrium.

Notice something else; the magnitude of the kinetic or potential energy involved in this oscillation doesn't depend at all on the mass of the object on which is oscillating! It depends only on the spring constant k and the amplitude of oscillation A .

At first this is surprising. Wouldn't a brick oscillating on a spring carry more energy than a tennis ball on the same spring? The solution to this paradox is clear when you think about the details. First, the energy you put into the system shows up at one point entirely in the stretching of the spring. At this point, all the energy is there in the spring. This stored spring energy will then be converted to kinetic energy in the object, but the amount of energy depends only on how much you stretch the spring. If you have a brick on the spring, it will convert all this into potential energy, but can do this while traveling more slowly. If it's

a tennis ball, it can still take all the energy but will have to travel faster. So the energy in the oscillating system depends only on the amplitude and spring constant, and not on the mass of the object which is moving.

Here's an example of the trading of energy in such an oscillation:



What is the total energy as a function of time during these oscillations?

$$KE + PE = \frac{1}{2}kA^2 \sin^2(\omega t) + \frac{1}{2}kA^2 \cos^2(\omega t)$$

$$KE + PE = \frac{1}{2}kA^2 (\sin^2(\omega t) + \cos^2(\omega t)) = \frac{1}{2}kA^2$$

In this last line we used the trigonometric identity $\sin^2 A + \cos^2 A = 1$. This equation says that the energy stored in an oscillator is *constant*. That's not too surprising, as we imagined that no other forces act.

This system, a simple harmonic oscillator, is one of the most important models in physics. It has very simple properties. The position, velocity, and acceleration of the object which is oscillating are all sinusoidal and related in a very simple way. The frequency of oscillation is governed by a balance between inertia and the strength of the restoring force. The energy associated with the oscillations is constant, though it is traded back and forth between kinetic and potential forms. The total amount of energy in the oscillations depends on the strength of the restoring force and the amplitude.

Later in this chapter, we will model a few important complications to this simplest system. But first we will generalize our discussion of potential energies a bit and see why the linear restoring force and SHM are such important models to develop.

Small oscillations are always SHM

We have seen before that potential energy minima will be equilibrium points around which oscillations will occur. Is the model of SHM motion induced by a linear restoring force always relevant for these cases? Let's imagine a completely general potential function which we can describe as a function $U(x)$. For this purpose it might be useful to think about gravitational potential for a bicyclist riding up and down hills. The bottoms of hills are all minima where she could rest at stable equilibrium.

Let's imagine our model potential energy function $U(x)$ has a minimum at the point x_0 . We can always approximate this function near any point by using a Taylor series. If we do this around the point x_0 we get

$$U(x) \approx U(x_0) + \left. \frac{dU}{dx} \right|_{x_0} (x - x_0) + \frac{1}{2} \left. \frac{d^2U}{dx^2} \right|_{x_0} (x - x_0)^2 + \frac{1}{3!} \left. \frac{d^3U}{dx^3} \right|_{x_0} (x - x_0)^3 + \dots$$

This is something we can always do, around any point, and if we keep x close to x_0 , the approximation can be very accurate, even if we keep only the first few terms. Now remember that in this case, the point x_0 is a minimum of this potential energy function $U(x)$. As a result, we know:

$$\left. \frac{dU}{dx} \right|_{x_0} = 0$$

And we can rewrite the full Taylor series approximation as:

$$U(x) \approx U(x_0) + \frac{1}{2} \left. \frac{d^2U}{dx^2} \right|_{x_0} (x - x_0)^2 \dots$$

Notice that this potential function has the same form as the potential function for the simple harmonic oscillator. In this version, the 'spring constant' k is given by

$$k = \left. \frac{d^2U}{dx^2} \right|_{x_0}$$

As long as the oscillations are small enough, the potential energy around **any** equilibrium position acts just like the potential energy of a simple harmonic oscillator. This is why, even though it is a simple model, SHM is such an important example to understand.

An example: a model intermolecular force

As an example, let's consider a model for the potential energy associated with two molecules which might bond with one another. It is typical for such a potential energy to include two different regions. When the molecules are very far apart the potential energy is independent of distance, and there is no force between the two. In fact, it makes substantial sense to *define* the potential energy when they are far

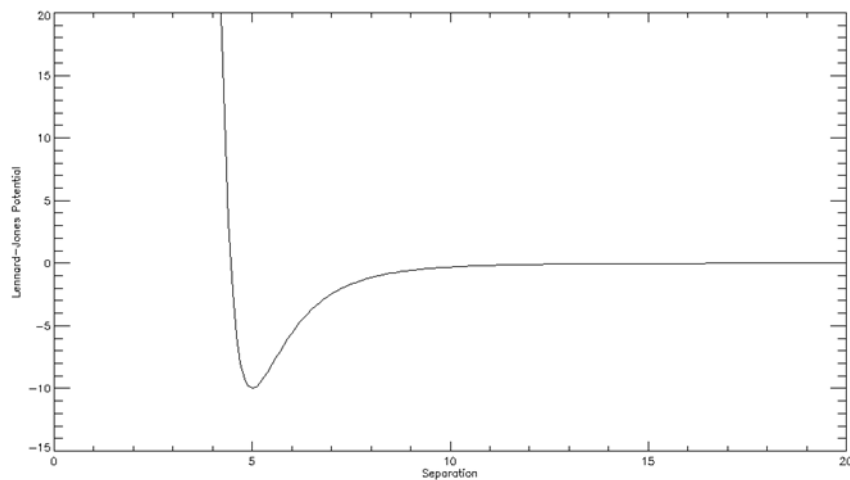
apart to be zero. If we do this, we will usually find that the potential energy of the two initially declines as they come closer together. Because of this declining potential energy, the two molecules will be pulled together by a ‘restoring force’.

When the two molecules get very close together, they begin to really run into one another, essentially trying to occupy the same space. At these very small separations, the potential energy of the two molecules can become very large and positive. As a result, there is typically a place between these two where the potential energy of the two molecules is minimized, and that’s the separation at which the two will typically bond together.

It’s not surprising that intermolecular forces and potential energies are complex. There is a mathematical model called the Lennard-Jones potential which has the features described above. It can be written:

$$U(r) = A \left[\left(\frac{r_0}{r} \right)^{12} - 2 \left(\frac{r_0}{r} \right)^6 \right]$$

Here is a plot of this potential function with the parameters chosen to be $A = 10$ and $r_0 = 5$:



Where is the minimum of this potential? It happens where the derivative of the potential with respect to r goes to zero. We calculate this as:

$$\frac{dU(r)}{dr} = A \left[-12 \left(\frac{r_0}{r} \right)^{11} + 12 \left(\frac{r_0}{r} \right)^5 \right] = 0$$

This condition is met when $r = r_0$, so r_0 is the equilibrium position. If we create the Taylor expansion of this function around this point we have:

$$U(r) \approx U(r_0) + \frac{1}{2} \left. \frac{d^2U}{dx^2} \right|_{r_0} (r - r_0)^2 \dots$$

$$U(r) \approx -10 + \frac{1}{2}(72)(r - r_0)^2 \dots$$

As we suggested above, this is a potential energy which takes the form of the potential that leads to simple harmonic motion. In this case the ‘spring constant’ k would take the value 72. The analysis applied here is in fact completely general. Small oscillations around any position of stable equilibrium will be well approximated by SHM with a spring constant given by the second derivative of the potential function at the equilibrium point.

12.4 First variation: damped oscillations

The first important addition to our simple model of SHM involves introducing a quick dose of reality. No real systems oscillate forever. They all eventually stop. This happens because friction always resists the motion, gradually draining away the energy which was so happily trading between kinetic and potential forms.

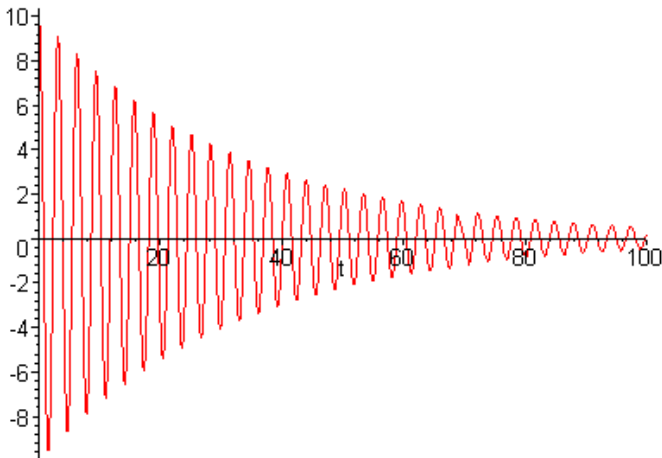
If there is friction slowing motion in an oscillator, it may have either a very small effect on the system (simply making the oscillations very gradually fade away), or it may have a very large effect, preventing oscillations completely or even preventing the restoring force to gradually drag the object back to equilibrium. Before doing any mathematical analysis, let’s consider what these different limits might look like.

Weak damping

Imagine first the case of friction which is relatively weak. Weak relative to what? If, during a typical oscillation, the size of the frictional force is always small compared to the average size of the restoring force, the friction is weak. Another way to say this is to insist that the amount of energy drained from the system during one cycle should be small compared to the overall energy of the system.

Systems like this are called “underdamped”. They will oscillate many times before coming to rest, each time losing some small fraction of their energy to friction. So in these systems, we expect to see the amplitude of oscillations gradually decrease.

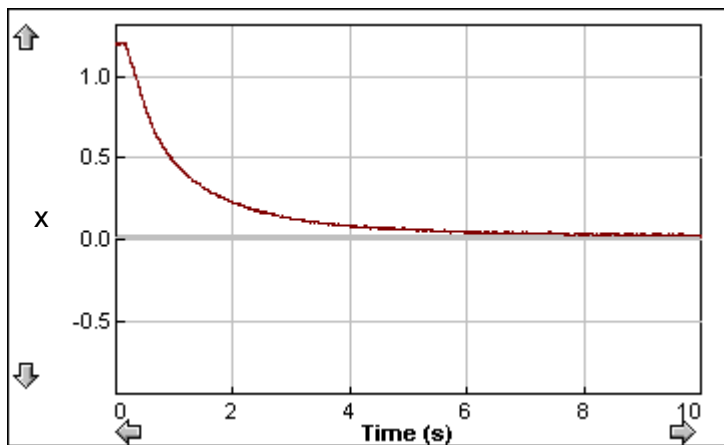
Does this damping affect the *frequency* of oscillation? In fact it does. Think about the first cycle. Without damping, the restoring force would accelerate the mass back toward equilibrium and it would take some time to get there. With damping, there is a resistive force always working against the restoring force. This resistive force slows the motion, making the frequency of oscillation lower.



This is an example of underdamped motion. Although the amplitude of each oscillation is somewhat less than the one before, there are many oscillations before the system finally comes to rest.

Overdamped motion

If the frictional force is really large, the system will never oscillate at all. Depending on the nature of the friction, it may never even move at all. Here's what happens. You pull the object away from equilibrium and let it go. The restoring force tries to pull it back to equilibrium. If the friction is velocity dependent (as it is for fluid friction) the object begins to speed up, but it may be that the frictional force quickly becomes large enough to completely balance the restoring force. When this happens, the object moves along at constant speed; in a motion completely analogous to terminal velocity. It's a little different from that case though, because the restoring force *isn't constant*. As the object gets closer to equilibrium, the restoring force gets smaller. The frictional force needed to balance that weaker restoring force is smaller, and the 'terminal velocity' is smaller. This balance gives rise to a gradual, smooth return to equilibrium, without any overshooting or oscillations. You pull it away from equilibrium, and it just gets smoothly dragged back.



This is an example of overdamped motion. In this case the frictional force associated with the damping is so large that it prevents any oscillations from ever happening at all. Instead, the object is just slowly dragged back to equilibrium by the restoring force.

Notice how in this case the slope of this position time graph starts out steep (with large negative velocity) and then becomes more and more shallow (with velocity decreasing). This shows how a linear restoring force differs from the terminal velocity case we considered before.

What happens if the friction force is not velocity dependent, but is something like the sliding friction of contact between solids? In this case, we begin by comparing the restoring force to the maximum static friction which could exist. If the maximum static friction is large enough, the object won't move at all. If not, then it will begin to slide toward equilibrium, gaining energy from the work done by the spring and losing energy due to the work done by friction.

One specific model for damped oscillators

How can we include this in our mathematical model for this motion? We have to add a new friction term to the force equation and see what motion this implies. As we have seen, friction is complex, and may be described by a wide variety of different phenomenological force laws. Most of these generate equations of motion which lack analytic solutions. But there is one which does have simple solutions. We will examine this one in detail, recalling that it is just an example, and will not apply in detail to every case.

If the friction which impedes the motion of our oscillator is proportional to the speed v , and always in a direction opposite to it, we could describe it with the equation:

$$F_{\text{friction}} = -bv$$

Where b is just some parameter describing how large this friction is. This might be the 'small-slow' friction associated with fluid motion which we've discussed before. For a small sphere moving slowly in a fluid this would be: $F_f = -12\pi\eta Dv$ where D is the diameter of the sphere and η is the viscosity of the fluid. For this kind of friction, we'd have $b_{\text{eff}} = 12\pi\eta D$. If we happen to have friction of this form, we can write the total equation of motion of the object as:

$$\sum F = -kx - bv = ma$$

To solve this we can write it as a 'differential equation'. Don't worry if you don't know how to do this, I just want to show you how it works in general.

$$m \frac{d^2x}{dt^2} = -kx - b \frac{dx}{dt}$$

Rearranging we find:

$$\frac{d^2x}{dt^2} + \left(\frac{b}{m}\right) \frac{dx}{dt} + \left(\frac{k}{m}\right) = 0$$

Recognizing that k/m is the quantity which, in the undamped oscillator, would be the square of the angular frequency of oscillation ω_0^2 , we can write:

$$\frac{d^2x}{dt^2} + \left(\frac{b}{m}\right)\frac{dx}{dt} + \omega_0^2 = 0$$

A “solution” to this linear differential equation is just some function $x(t)$ which obeys this equation. The general form for this solution can be written:

$$x(t) = Ae^{-\frac{b}{2m}t} \cos(\omega't + \phi)$$

Where

$$\omega' = \sqrt{\omega_0^2 - \left(\frac{b}{2m}\right)^2}$$

How do you find this solution? It turns out there are a variety of techniques for finding the solutions to differential equations. You'll learn about them when you take more advanced math courses. For now, we will simply demonstrate that this solution is valid. To do this, we take the proposed solution $x(t)$ and put it into the equation of motion. If it is a correct solution, the equation will be satisfied. To begin, we find the appropriate derivatives of $x(t)$:

$$\frac{dx}{dt} = -\left(\frac{b}{2m}\right)Ae^{-\frac{b}{2m}t} \cos(\omega't + \phi) - A\omega'e^{-\frac{b}{2m}t} \sin(\omega't + \phi)$$

$$\begin{aligned} \frac{d^2x}{dt^2} = & \left(\frac{b}{2m}\right)^2 Ae^{-\frac{b}{2m}t} \cos(\omega't + \phi) + \left(\frac{b}{2m}\right)\omega' Ae^{-\frac{b}{2m}t} \sin(\omega't + \phi) + \\ & \left(\frac{b}{2m}\right)\omega' Ae^{-\frac{b}{2m}t} \sin(\omega't + \phi) - A\omega'^2 e^{-\frac{b}{2m}t} \cos(\omega't + \phi) \end{aligned}$$

Now put these into the differential equation above:

$$\begin{aligned} & \left(\frac{b}{2m}\right)^2 Ae^{-\frac{b}{2m}t} \cos(\omega't + \phi) + \left(\frac{b}{m}\right)\omega' Ae^{-\frac{b}{2m}t} \sin(\omega't + \phi) - A\omega'^2 e^{-\frac{b}{2m}t} \cos(\omega't + \phi) + \\ & \left(\frac{b}{m}\right)\left(-\left(\frac{b}{2m}\right)Ae^{-\frac{b}{2m}t} \cos(\omega't + \phi) - A\omega'e^{-\frac{b}{2m}t} \sin(\omega't + \phi)\right) + \omega_0^2 = 0 \end{aligned}$$

Simplifying, we find:

$$\left[\omega_0^2 - \left(\frac{b}{2m} \right)^2 - \omega'^2 \right] A e^{-\frac{b}{2m}t} \cos(\omega't + \phi) = 0$$

This equation will be satisfied so long as:

$$\omega' = \sqrt{\omega_0^2 - \left(\frac{b}{2m} \right)^2}$$

Which is the condition on this parameter ω' which we specified above. So this solution works. This function $x(t)$ does describe the motion of an object under the influence of the force of the form

$$F_{\text{total}} = -kx - bv.$$

These solutions have three different “regimes”, three different general behaviors which depend on the specific choices for the parameters b , m , and k :

$$\text{Underdamped:} \quad \frac{k}{m} > \left(\frac{b}{2m} \right)^2 \quad \text{or} \quad \omega_0^2 > \left(\frac{b}{2m} \right)^2$$

$$\text{Overdamped} \quad \frac{k}{m} < \left(\frac{b}{2m} \right)^2 \quad \text{or} \quad \omega_0^2 < \left(\frac{b}{2m} \right)^2$$

$$\text{Critically damped} \quad \frac{k}{m} = \left(\frac{b}{2m} \right)^2 \quad \text{or} \quad \omega_0^2 = \left(\frac{b}{2m} \right)^2$$

What will happen in the oscillator depends on the balance of damping and restoring force in the system. If damping is weak, it may oscillate for a long time. If damping is strong, it may not oscillate at all. Let's examine in more detail the solution in each of these three cases.

First consider the overdamped case. Since in this case $\omega_0 > b/2m$:

$$\omega' = \sqrt{\omega_0^2 - \left(\frac{b}{2m} \right)^2} = \sqrt{\text{negative number}}$$

Here we have ω' being the square root of a negative number! So we will have

$$\omega' = \pm i \sqrt{\left(\frac{b}{2m} \right)^2 - \omega_0^2} = \pm iS \quad \text{where} \quad S = \sqrt{\left(\frac{b}{2m} \right)^2 - \omega_0^2}$$

and since (OK this is something you may not have learned yet, but is a basic result of complex analysis)...

$$\cos(\theta) = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$$

we can write:

$$\cos(\omega't + \phi) = \frac{1}{2}(e^{i(\omega't + \phi)} + e^{-i(\omega't + \phi)}) = \frac{1}{2}(e^{-St} e^{i\phi} + e^{St} e^{-i\phi})$$

This makes the whole solution:

$$x(t) = A'e^{\left(\frac{b}{2m} + S\right)t} + B'e^{\left(\frac{b}{2m} - S\right)t} = A'e^{-\mu t} + B'e^{-\kappa t}$$

Where we have absorbed the addition factors $e^{i\phi}$ and $e^{-i\phi}$ into the new constants A' and B' , and define the two new decay constants as:

$$\mu = \left[\frac{b}{2m} - \sqrt{\left(\frac{b}{2m}\right)^2 - \omega_0^2} \right] \quad \text{and} \quad \kappa = \left[\frac{b}{2m} + \sqrt{\left(\frac{b}{2m}\right)^2 - \omega_0^2} \right]$$

Putting this rather complicated thing into words, the overdamped solution is a decaying double exponential function. When an oscillator is overdamped in this way, the displacement just decays away to zero.

Now let's consider the critically damped case. Here we have:

$$\omega' = \sqrt{\omega_0^2 - \left(\frac{b}{2m}\right)^2} = 0$$

so the full solution is just:

$$x(t) = Ae^{-\left(\frac{b}{2m}\right)t} \cos(\phi)$$

This one is just a pure single exponential decay. It doesn't oscillate at all.

In the third, underdamped case, ω' is a real, positive number, and the solution is the full:

$$x(t) = Ae^{-\left(\frac{b}{2m}\right)t} \cos(\omega't + \phi)$$

There are two things to note about this solution. It basically has two parts; an oscillatory cosine function and a time varying amplitude. The amplitude of the oscillation decreases exponentially with time. This is governed by the term

$$Ae^{-\frac{b}{2m}t}$$

in front of the oscillatory cosine function. This coefficient $b/2m$ is sometimes rewritten as $1/\tau$, where τ is the ‘decay time’ of the system. Making this substitution, we can write the amplitude in this form:

$$Ae^{-\frac{t}{\tau}}$$

From this, you can see that the amplitude of oscillation falls off by a factor of e^{-1} when $t = \tau$.

In addition to the declining amplitude the oscillations of the damped oscillator take place at a lower frequency than they would in the undamped ($b = 0$) case. As the damping becomes larger, this shift in frequency gets larger and larger, until eventually the frequency goes to zero and no oscillations ever occur.

As $b/2m$ gets closer to ω_0 , the frequency of oscillation falls, until at the limit $\omega_0 = b/2m$, we have zero frequency and an infinite period. So remember that *two* things happen with damped oscillations:

- The amplitude drops exponentially with time
- The frequency is shifted lower

Now this set of mathematics give specific predictions for the particular case where friction is proportional to velocity. But these basic facts, that in a damped oscillator the amplitude will gradually drop to zero and the frequency will be shifted lower, are true for any kind of damping. That’s why we introduced the basic ideas before doing this one mathematical example.

12.5 Natural frequencies of oscillation

We have seen that oscillators, if disturbed, will oscillate with a frequency that is determined by a balance of the strength of the restoring force and the inertia of the system. In the derivation above, we have added to that an adjustment based on the strength of the damping experienced by the system. But it is still the case that, for any oscillator, there is a ‘natural frequency’ at which it will oscillate if you disturb it. How do you find this natural frequency? This is simple, just disturb it and watch to see what frequency it oscillates with. For example, if you want to know the natural frequency of a playground swing is, you just get on and push off. You’ll oscillate back and forth and can measure the natural frequency.

Driven oscillations and resonance

As we mentioned at the start, it is *often* important to consider what will happen to an oscillator if you hit it with a periodic force. Most often this happens because two oscillators are sitting next to one another (like two atoms in a solid) and once one starts oscillating it repeatedly strikes the next.

How might we model this mathematically? We could start with the equation we wrote for a damped oscillator and now include a new ‘driving’ force in the equation.

$$\sum F = ma = m \frac{d^2x}{dt^2} = -kx - b \frac{dx}{dt} + F_{\text{driving}}(t)$$

This is typically rewritten as:

$$\frac{d^2x}{dt^2} + \left(\frac{b}{m}\right)\frac{dx}{dt} + \omega_0^2 = \frac{F_{\text{driving}}(t)}{m}$$

What form might we expect to have for this driving force? The answer of course depends on the situation. We might drive the oscillator by simply shoving it once, then letting it oscillate freely. You can imagine what would happen if you did this to your friend who was sitting on a swing. You would give one hard push and then let her swing. She would oscillate back and forth for a while, then gradually come to a halt.

Or we might drive the oscillator by pushing in a regular, periodic fashion, something like:

$$F_{\text{driving}}(t) = F_0 \cos(\omega_{\text{driving}} t)$$

Solving the equations of motion with a driving force like this is a complicated business; best left to a mathematical methods class. The general solution is variable in time, oscillating wildly and not just sinusoidally. But there is a simple, physically motivated prediction we can make. If we wait long enough, the system will eventually settle down and oscillate with the frequency of the driving force ω_{driving} .

Why is this? If the system starts out oscillating with some frequency that's *not* equal to the driving frequency, this part of the oscillation will eventually die out due to damping. Oscillations at other frequencies must die out because there is nothing in the system putting new energy into them. Without energy going in, damping will remove whatever they start with. What won't die out are those oscillations which happen at the frequency of the driving force, because these oscillations are continually restored by the driving force..

So the 'long time' solution will be a final oscillation with the frequency of the driving force. How long it takes to settle down into this mode depends on the size of the damping. If the damping is large, this will happen quickly. If the damping is small, this may take a long time. But we will always end up with oscillations at angular frequency ω_{driving} ; a fairly simple and perhaps not too interesting solution. What *is* interesting is the surprising relation between the amplitude of the final oscillation, the frequency of the driving force ω_{driving} , the natural frequency of the undriven oscillator ω' , and the size of the damping b . If the driving frequency is close to the natural frequency, the eventual amplitude of the oscillations will become large. If the driving frequency is far from the natural frequency, the eventual amplitude of the oscillations will be small.

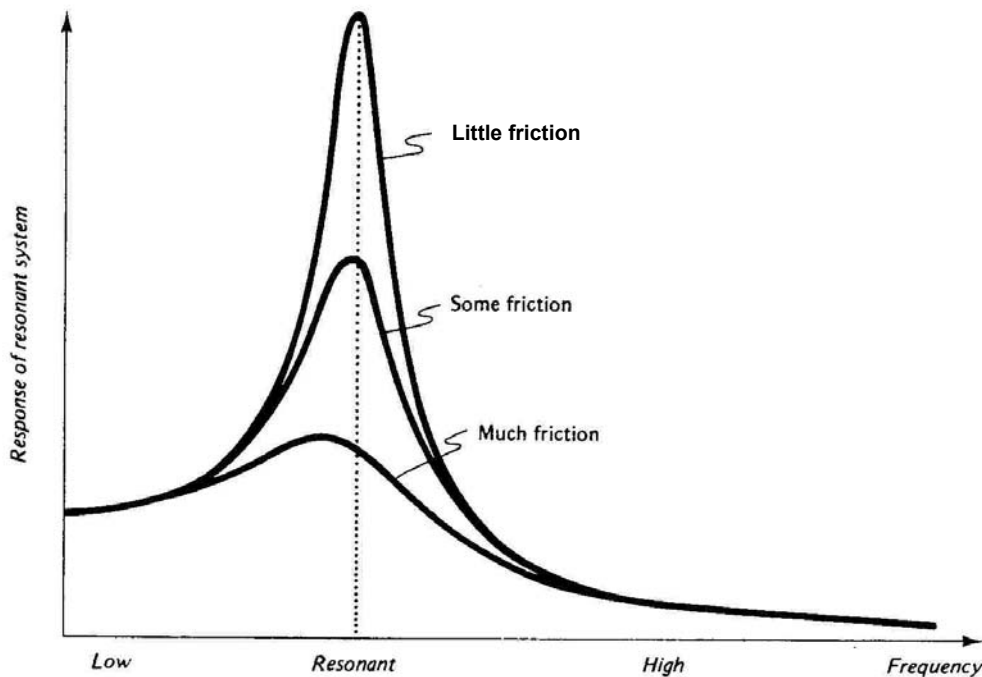
You are probably familiar with this from your own playground experience. If you go to push your friend Mary on a swing, you know that to build up a nice, large amplitude motion, you have to push at the same frequency with which Mary would naturally swing. If, for example, you push her back and forth really quickly, like five times a second, she will shake back and forth, but she'll never build up any large amplitude. Likewise, if you push her back and forth very slowly, say once every 30 seconds, again, she will build up no large amplitude. But if you push her at just the right rate, at just the natural frequency of the swing, large amplitudes will quickly appear.

How do you know what frequency to push with, and moreover, how do children know this in kindergarten? It's easy. You want to push with the natural frequency of the swing. To find this, you just

give a shove and watch Mary swing. Once you see this natural “free” oscillation, you know how often to push.

How large an oscillation will we get? If we drive the system at its "natural frequency" ω' (from above) then it oscillates with large amplitude; how large depends on the damping. If we drive it at a frequency different from this, it oscillates at the new frequency, but with smaller amplitude; essentially the oscillation always fights the driving force. The relation between the frequency of the driving force and the amplitude of the final oscillation is called the resonance curve, which shows what final amplitude you'd get if you drove the oscillator at each driving frequency.

Not surprisingly, if you have damping, the maximum amplitude you can reach while driving the oscillations is less. In fact the shape of the ‘resonance curve’ changes in several ways as the damping increases. First, the peak frequency (the natural frequency) shifts lower. This is just the effect we saw already ($\omega'^2 = \omega_0^2 - b^2/2m^2$). Second, the maximum amplitude of oscillation decreases. Since energy is being drained out more quickly, you just don't build up as large an amplitude. Finally, the overall width of the observed resonance curve increases with damping.



So for driven systems, you get oscillations at the driving frequency, and their amplitude depends on the relation of the driving frequency to the "natural" frequency of the system. If there is a lot of damping, oscillations will never be large. If there is little damping, and you drive the system at the right frequency, the amplitude can be very large.

Familiar examples of resonance

The playground swing is the most obvious and familiar. There are many similar examples though. If your Frisbee is stuck in a small tree you might shove the tree repeatedly, trying to shake it loose. In doing this, you will, without thinking about it, note the natural period of oscillation of the tree and shove in synch with this. Working in this way allows you to drive the oscillations of the tree to large amplitude, and hopefully knocks free your Frisbee. If your car is stuck in a ditch and your friends help you to push it out, you may 'rock' it back and forth, shoving forward at just the frequency with which it would oscillate.

Another familiar example involves walking with a cup of coffee. You can measure the natural frequency of the system ω' by disturbing it, just giving it a nudge. If you walk at that frequency, you will drive these oscillations in resonance, and the coffee will start sloshing and may spill all over you.

A Quick Summary of Some Important Relations

Basic relations for harmonic motion:

When something happens regularly, it has a period, frequency, and amplitude. Period and frequency are related in this simple way:

$$\text{Period} = \frac{1}{\text{Frequency}} \quad \text{or} \quad T = \frac{1}{f}$$

Simple harmonic motion:

Whenever the force restoring an object to equilibrium is proportional to the distance from equilibrium, oscillations can be described as simple harmonic motion. When the displacement from equilibrium is described as $x(t)$, this motion is governed by these relations:

$$x(t) = A \sin(\omega t) \quad v(t) = -\omega A \cos(\omega t) \quad a(t) = -\omega^2 A \sin(\omega t) = -\omega^2 x(t)$$
$$\omega = 2\pi f = \frac{2\pi}{T}$$

Remember that these relations tell you the maximum displacement, velocity, and acceleration during these oscillations.

Examples of SHM:

Mass on a spring:

$$F_{\text{restoring}} = -kx$$
$$\omega = \sqrt{\frac{k}{m}}$$

Simple pendulum:

$$F_{\text{restoring}} \propto mg$$
$$\omega = \sqrt{\frac{g}{L}}$$

All oscillations around equilibrium are well modeled as SHM for small enough amplitudes.

Energy in SHM:

In SHM energy trades back and forth between kinetic and potential. For a mass on a spring the total energy is $E_{\text{total}} = \frac{1}{2} kA^2$.

Damped oscillations in general:

When friction is present in the motion of an oscillator, it will gradually drain away the energy present in the oscillations. The importance of this depends on the relative size of the frictional forces and the restoring force. When damping is small, the system is called ‘underdamped’, and will oscillate many times before coming to rest. When damping is large, the system is called ‘overdamped’, and will not oscillate at all – it will slowly return to equilibrium. At the balance between these two the system is called ‘critically damped’, and returns to equilibrium, without oscillating, as quickly as possible.

Damped oscillations with small-slow fluid friction:

In the special case an oscillating object with mass m experiences fluid friction of the small-slow form, the shift in natural frequency is given by:

$$F_{\text{restoring}} = -kx$$
$$F_{\text{fluid friction}}^{\text{small-slow}} = -bv$$
$$\omega_{\text{new}} = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2}$$

For this special case, the three possible outcomes for damped oscillators can be specifically defined.

Damped and driven oscillators and resonance:

Each oscillator has a ‘natural frequency’ at which it will oscillate if disturbed. When any kind of damping is present, it slows oscillations and lowers this natural frequency. When any oscillator feels a force which drives it periodically, the result will depend on whether the driving frequency is close to or far from the natural frequency. Driving an oscillator at its natural frequency leads to large amplitude, possibly dangerous, oscillations. When damping is small, resonant oscillations can be large, but only build up when the driving frequency is close to the natural frequency. When damping is large, resonant oscillations will never become large.

13. Energy at the atomic level: heat and thermal motion

- 1) Energy contained in a material: temperature and thermal energy
 - i. Conservation of energy
 - ii. Temperature and heat energy
- 2) An ideal gas is simplest
 - i. Pressure and the equation of state
 - ii. Moles and numbers of atoms
 - iii. The ideal gas law
- 3) Origins of the ideal gas law
 - i. Atoms and their motions
 - ii. What is temperature really?
- 4) Why things happen the way they do: the fundamental assumption of statistical physics
 - i. Predictable outcomes from random motion: irreversibility
 - ii. Irreversibility and chance
 - iii. Using this in statistical physics
- 5) Entropy and statistics

Physics for the Life Sciences: Chapter 13

In this chapter we will build on the introduction to energy and motion we have been constructing. So far our notions of energy have been macroscopic, contained in the tangible motions of sizable things, as well as in their interactions with one another. Macroscopic motion we account for as kinetic energy. The energy associated with interactions we account for as potential energy. At this point we're ready to introduce another, apparently very different form of energy; thermal energy.

Thermal energy in our surroundings is obvious and active. Its warmth makes a spring day delightful; its absence in winter stops life almost completely. When it is intense, as in a fire, it seems practically alive and illuminates its surroundings. Thermal energy is clearly an agent of change in the world. Heat melts snow and ice, incubates eggs, and gives our French Fries that lovely crispy exterior. It was not until well into the 19th century that physicists like James Joule and Lord Kelvin could understand with any confidence how thermal energy and the macroscopic mechanical energy we have been discussing could be converted, one into the other.

The connection was not obvious for two reasons. First, they knew nothing of atoms. So they had no idea what changed in an object when it went from hot to cold. An atomic understanding of matter, and its connection to thermal energy, would not emerge until statistical physics matured with the work of James Clerk Maxwell and Ludwig Boltzmann in 1860s. The second challenge in relating thermal and mechanical energy is a quantitative one. The amount of energy required to alter the temperature of a substantial object like a book is very large. So, for example, when a book slides across a table and comes to rest at the other end, the energy which friction converts to heat has a small, hardly perceptible, affect on the temperature of the book. To measure these small changes in temperature required exquisite experimental skill. In some ways modern experimental physics and the study of these thermal phenomena were born together.

In this chapter, we will show how thermal energy is present in the kinetic and potential energies of the atoms which make up a substance. In so doing, we will develop a clear sense of what thermal energy is and of why heat flows in the ways it does. In this explanation we find the germ of the most profound contribution of statistical physics; an explanation for why, among a limitless variety of allowed

possibilities, only a small subset of outcomes ever occurs. This is a very deep idea, explaining, among other things, why irreversible processes exist, decay is inevitable, and entropy always increases.

13.1 Energy and its forms

The idea of energy was not part of Newton's physics. There are many reasons for this, not least that the energy concept is subtle and complicated. The first form of energy to be accounted for in the framework of Newtonian mechanics was kinetic energy. Thomas Young introduced KE as an important quantity to track in physics, and showed that KE is changed when work is done.

$$KE = 1/2mv^2$$

$$W = \int \mathbf{F} \cdot d\mathbf{s} = \Delta KE$$

Later in the 19th century, William Rankine introduced the idea of potential energy as a way of accounting for the interaction energy associated with the configuration of a system. We have explored the energy associated with the gravitational pull of the Earth and the energy associated with elastic deformation. These are our first two forms of energy: kinetic and potential.

It turns out there are many others as well. The fact that many forces can do work, changing the kinetic energy of an object, without being "conservative" forces associated with potential energies is strong evidence for this. Perhaps it is best to consider an example. Imagine that you slide a book across a table. It starts out with kinetic energy, then gradually slides to a halt as the force of friction does negative work on it. What happens to the energy the book had? It is not lost, or "used up" by the friction force. Instead, it is converted to another form. The energy which the book started out with is instead used to raise the temperature of the book and the table.

To understand this process, we need to figure out how to include heat, which is clearly a form of energy, in the tidy framework we established for studying kinetic and potential energy.

Conservation of Energy

The conservation of energy, the notion that it can neither be created nor destroyed, was originally simply an observation. Preserving this observation initially required a kind of tricky bookkeeping. In the little example above, we preserved energy conservation by asserting that the kinetic energy originally present was somehow turned into heat energy. That kinetic energy is gone, sure, but look, here it is in a new form we weren't worrying about before. That might seem like sleight of hand, like a rhetorical trick; if it disappears we just give it a new name. Absent any more fundamental reason for believing in the conservation of energy, this would certainly be a valid criticism.

Richard Feynman, an influential 20th century American physicist, described the situation this way:

"There is a fact, or if you wish, a law, governing natural phenomena that are known to date. There is no known exception to this law—it is exact so far we know. The law is called conservation of energy; it states that there is a certain quantity, which we call energy that does not change in manifold changes which nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says that there is a numerical quantity, which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a

strange fact that we can calculate some number, and when we finish watching nature go through her tricks and calculate the number again, it is the same..."

If energy conservation were based only on observation, if we only thought it was conserved because we'd always seen it conserved, you might hope to violate energy conservation. In this case, you might be able to create energy, to make it without cost. Obviously a world with this kind of free energy would be marvelous, and some people continue to seek it. In fact, there are people who will sell you stock in their free energy company online.

Fortunately, (unfortunately?) we also have a deeper understanding of the origin of energy conservation, one which provides a convincing theoretical foundation for the conservation we observe. It was discovered in the first part of the 20th century by a remarkable mathematician named Emmy Noether. She was able to connect conservation laws (like the conservation of energy and momentum) in a rigorous way to symmetry principles. What does this mean?

First, what's a symmetry in a physical law? Saying that a physical law has a symmetry just means that the law is unchanged when you alter it in some way. For example, the laws of physics have a spatial symmetry. They are the same whether you study them in Ann Arbor or Uttar Pradesh. In fact, they are known to be the same both here and in the most distant parts of the universe. As far as we know, the laws of physics also possess a time symmetry. They are the same now as they were in Newton's time. They seem not to have changed since the big bang, and we expect them to be the same tomorrow.

Noether proved that there *must* be a conserved quantity associated with each symmetry present in the laws of physics. Any violation of the symmetry would require a violation of the conservation law, and any violation of the conservation law would require a violation of the symmetry.

As it happens, the conservation of energy derives from the time symmetry of the laws of physics. Saying that energy is conserved is equivalent to saying that the laws of physics we learn today were correct yesterday, and will be the same tomorrow. Now this seems like a pretty fundamental statement. Without it, the rules which the universe obeys would be different from day to day, a situation which would make a scientific approach to the world rather less useful. In a certain sense, if there's any point in attempting to understand physical laws, then energy is going to be conserved.

Should this make you much more confident in energy conservation? Perhaps not, but it should help you to understand why a scientist responds to claims that energy is not conserved with deep skepticism. If such a claim were true, the implications would be quite staggering. This is one of the key reasons why you should be *extremely* skeptical if someone tries to sell you free energy.

Temperature and heat energy

We know from experience that when friction (a "non-conservative" force) acts, things heat up. For example, you probably rub your hands together to warm them on a chilly morning. We're saying now that this "heating up" is due to the deposition of energy taken from motion by friction. It seems clear that heat energy is associated with temperature, but what exactly is the connection? To see this we need first to take a step back and talk about how temperature scales are defined.

There are two important temperature scales in use in science. Each is defined by fixing the temperature at which water freezes and boils, then dividing the temperature range between these into 100 equal intervals called degrees. The Celsius scale sets the freezing point to 0° C and the boiling point to 100° C. The Kelvin scale sets the freezing point to be 273.15 Kelvin and the boiling point to be 373.15 Kelvin. By convention, we say ‘degrees Centigrade’ when using the Celsius scale and simply ‘Kelvin’ when using the Kelvin scale.

Measurement of temperature relies on the fact that many properties of matter change in a fully repeatable way with temperature. To take just one of many examples, liquids and solids expand and contract as temperature changes. This is the basis of mercury thermometers. Once calibrated, changes like this can be used to build devices which reliably measure temperature. While expansion of any solid or liquid can be used to define a temperature scale, there are drawbacks to this approach. The scale you get doing this is very specific, and will be a somewhat different if you use different materials.

It turns out that a much more universal temperature scale can be defined using the properties of dilute gases. What do we mean by a “dilute” gas? Basically, we just need to make sure that the typical distance between particles in the gas is very large compared to the size of the particles. This condition is easily met by all the gases you usually encounter. Such dilute gases have remarkably simple, and similar, relations between physical properties and temperature. We will see why in a bit, but first let’s explore the essential phenomena.

13.2 A simple model for a gas

In an ordinary gas individual atoms or molecules move about freely. Usually they’re just flying through space, though occasionally they run into one another or into the walls of the vessel in which they’re contained. When they do, they bounce off one another elastically. In an “ideal” gas, there’s no interaction among the atoms at all. They never hit one another, though they do still hit the walls of the vessel. Each collision of an atom with the wall exerts a tiny, very brief force. Taken together, this enormous number of continuous, tiny impacts creates a steady average force pushing out against the walls. The force is spread nearly uniformly across the walls, and so is described as a pressure.

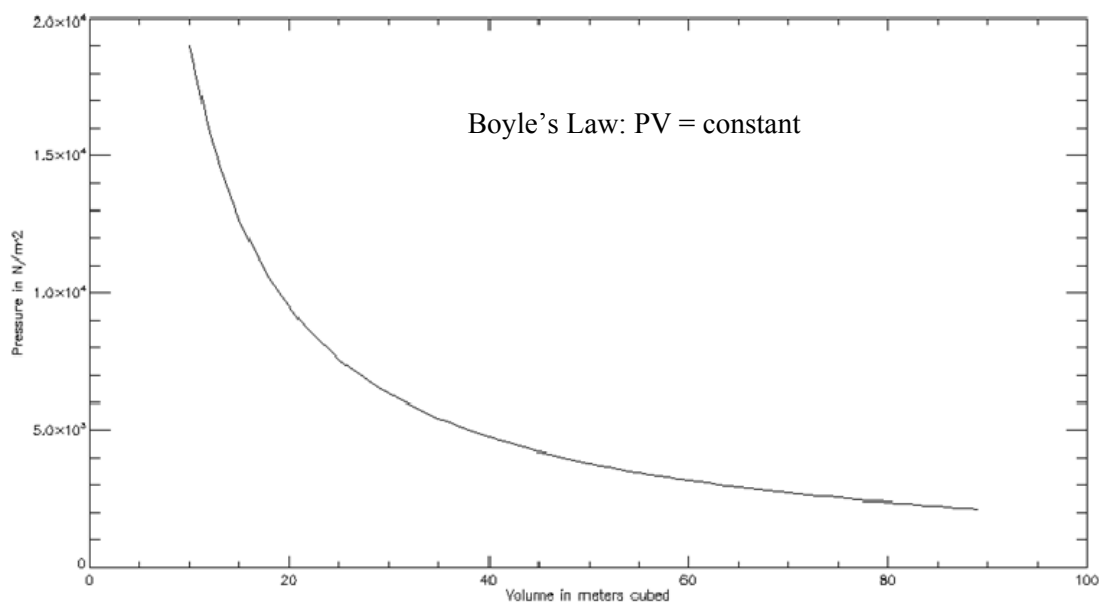
Pressure and the equation of state

Unlike solids and liquids, gases will expand freely to fill whatever contains them. In fact, they always push outward on the vessel which contains them, exerting a uniform *pressure* on the walls of the container. This pressure, a force per unit area, is measured in N/m², and is uniform throughout the gas.

Early in the 19th century careful experiments in England showed that, if you keep a gas at a fixed temperature and alter its volume, the resulting pressure obeys “Boyle’s Law”:

$$PV = \text{constant}$$

If you increase the volume, the pressure goes down. If you decrease the volume, all the time holding the temperature constant, the pressure goes up. Note that this means if you completely remove the pressure, taking away the walls of the vessel, the gas will expand to infinite volume. Only the presence of the walls prevents this from happening.



Further experiments later showed another connection. If you keep the pressure constant and vary the temperature (measured on the Celsius scale), the volume of the gas changes according to a rule called Charles's law:

$$V_{\text{gas}} \propto (T_{\text{Celsius}} + 273.15)$$

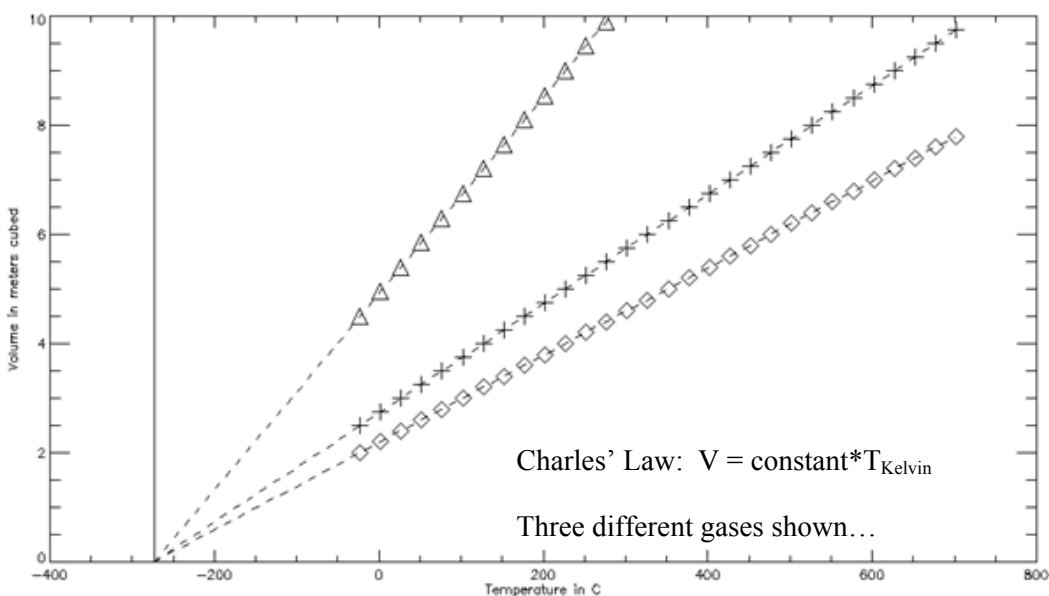
In words, the volume of any sample of gas was proportional to the temperature of the gas measured on the Celsius scale plus 273.15°C . The volume of any sample of dilute gas as a function of temperature is a line with an intercept at $T_{\text{Celsius}} = -273.15$. This discovery prompted the invention of a new temperature scale, called the "Kelvin" scale, which is defined by:

$$T_{\text{Kelvin}} = T_{\text{Celsius}} + 273.15$$

Using this temperature scale, Charles law can be written:

$$V_{\text{sample of gas}} = \text{constant} \times T_{\text{Kelvin}}$$

From experiments, it's possible to find these constants in Boyle's and Charles's Laws. Remarkably the constants you find are the same for *all* dilute gases, with only one requirement. To know what the constant is, you have to know how many atoms are in the gas you're looking at. It makes *no difference* what kind of atom or molecule they are; you only need to know how many.



Moles and numbers of atoms

How do you know how many atoms are in the gas you're looking at? It turns out the number of atoms in any reasonable amount of gas is *very* large, so you can't just count. Instead, we take the mass of each atom (or molecule if it's a molecular gas), and divide the total mass of the gas by the mass of each atom. Masses of atoms and molecules are measured in "atomic mass units", or amu. The scale for this is:

$$1 \text{ amu} = 1.66 \times 10^{-27} \text{ kg}$$

If your molecule weighs 28 amu (as does N_2 , the primary component of the atmosphere) and you have 100 gm of this gas, the number of atoms you have is:

$$\frac{0.1 \text{ kg}}{28 \times 1.67 \times 10^{-27} \text{ kg}} = 2.15 \times 10^{24} \text{ atoms}$$

which is a *lot* of atoms.

It is useful to define a particular number of atoms such that the total weight in **grams** is just equal to the particle weight in **amu**. For example, the Hydrogen molecule H_2 has a molecular mass of 2 amu, or about 3.32×10^{-27} kg. If we had 2 grams of H_2 we would have a total number of atoms:

$$0.002 \text{ kg} / 3.32 \times 10^{-27} \text{ kg/atom} = 6.024 \times 10^{23} \text{ atoms}$$

This particular number of atoms is called a "mole" of atoms. It's just a number, arbitrary really, that is useful for comparing the same number of atoms of different substances. It roughly reflects the number of atoms you might find in a typical modest sized sample of some material, and is often called Avogadro's

number. In a sense, it tells us the rough conversion between some ordinary thing and the number of atoms which make it up.

Using this definition, the mass of one mole of a substance is:

$$\begin{aligned}\text{Mass of one mole} &= 6.024 \times 10^{23} \text{ atoms} \times \text{number of amu} \times 1.67 \times 10^{-27} \text{ kg} \\ &= \text{number of amu} \times 10^{-3} \text{ kg}\end{aligned}$$

Using this definition, for a molecule with a molecular weight of 28 amu (like N₂), the mass of one mole is 28 grams.

The ideal gas law

As it turns out, you can combine Charles' and Boyle's Laws into a single "ideal gas law":

$$PV = nRT$$

Where P is pressure, V is volume, T is temperature in the Kelvin scale, n is the number of moles of gas in your sample, and R is a constant called the "universal gas constant" for rather obvious reasons. Experiments show that its value is:

$$R = 8.3 \frac{\text{Joules}}{\text{mole} \times \text{Kelvin}}$$

The units here are Joules per mole per Kelvin. Do these make sense? What are the units of PV ? Pressure is force / area (N/m²) and volume is m³, so PV has units of Nm, which is the unit of energy, the Joule. When you multiply R , with units of Joule / mole \times Kelvin, you get Joules. So these units work.

This law is quite accurate for all gases under ordinary conditions. It allows us to define a temperature scale in a more fundamental way. Take a fixed number of gas atoms (a number of moles n), put it in a vessel of volume V and measure the pressure P . Then the temperature T is given by:

$$T = \frac{PV}{nR}$$

This scale provides a material independent temperature scale. Use any gas you like (so long as it is dilute) and you'll get the same results. When talking about this, we'll sometimes say "atoms" or "molecules" or "particles". It doesn't really matter, as the result is independent of that detail, at least to the extent that the gas behaves ideally.

This gas temperature scale allows us to build a thermometer which is more general than a mercury or alcohol thermometer. We don't even have to know precisely what it's made of for it to work. This generality is very attractive to physicists, mostly because it hints at something significant going on, and provides a hint about where to seek the fundamental nature of temperature.

13.3 Origins of the ideal gas law: atoms and their motions

In our atomic picture, we imagine that a gas is made up of an enormous number of tiny particles, each traveling freely except when they collide elastically with one another or the wall. In this model, the space between particles is very large compared to their size, so they spend most of their time moving about freely, and only a small fraction of their time colliding with one another or the wall.

What happens when they strike the wall? Each time a particle hits the wall it bounces off elastically. The wall applies an impulse (a force for some period of time: $\vec{F}\Delta t$) to the particle, and the particle applies an equal and opposite impulse to the wall. When you average the effect of a huge number of impacts, the wall experiences some average force per unit area (a pressure). If you change the conditions, you can change this pressure. Increase the number of atoms in the vessel, and they will strike the walls more often, increasing the pressure. Increase the volume of the vessel and they will strike the wall *less* often, decreasing the pressure.

To see quantitatively how this works, consider a simple model with a single atom rattling around in a cubic box with edge length L . Each time it hits the wall on the right, there is a change in momentum $\Delta p = 2mv_x$ (because it bounces back). This happens once every time the particle crosses the box twice, a time $\Delta t = 2L/v_x$. Putting these together, we get an estimate of the average force exerted by one atom on the right hand wall:

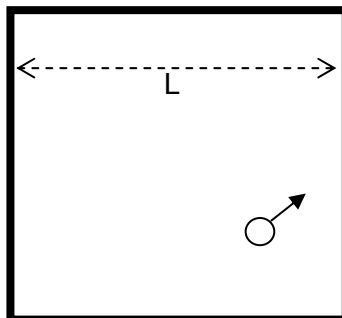
$$F_{\text{one atom}} = \frac{\Delta p}{\Delta t} = \frac{2mv_x}{(2L/v_x)} = \frac{mv_x^2}{L}$$

Multiply this by the number of atoms ($n_{\text{atoms}} = n_{\text{moles}} N_A$) in the box to get the total average force on the right hand wall, then divide by the area of the right hand wall to get the average pressure:

$$P = \frac{n_{\text{moles}} N_A F_{\text{one atom}}}{A} = n_{\text{moles}} N_A \frac{mv_x^2}{L \times A} = \frac{n_{\text{moles}} N_A}{V} mv_x^2$$

Rearranging this gives:

$$PV = n_{\text{moles}} N_A mv_x^2$$



This is an interesting result, because this mv_x^2 looks like part of the kinetic energy of the atoms. To see how it relates to the total kinetic energy, consider how the magnitude of v_x relates to the magnitude of v_y and v_z on average for these atoms.

If they're bouncing around over and over in this box, there's no reason for any one of the components (x, y, or z) to be any different from the others. So we expect on average to find:

$$\left(v_x^2\right)_{\text{average}} = \left(v_y^2\right)_{\text{average}} = \left(v_z^2\right)_{\text{average}}$$

This allows us to relate $\left(v_x^2\right)_{\text{average}}$ to the average total velocity squared:

$$\begin{aligned} \left(v_{\text{total}}^2\right)_{\text{average}} &= \left(v_x^2\right)_{\text{average}} + \left(v_y^2\right)_{\text{average}} + \left(v_z^2\right)_{\text{average}} = 3\left(v_x^2\right)_{\text{average}} \\ \left(v_x^2\right)_{\text{average}} &= \frac{\left(v_{\text{total}}^2\right)_{\text{average}}}{3} \end{aligned}$$

Plugging this into the above relation, and pulling out a factor of $\frac{1}{2}$ to explicitly include the kinetic energy, we find:

$$\begin{aligned} PV &= n_{\text{moles}} N_A \frac{\left(mv_{\text{total}}^2\right)}{3} = \frac{2}{3} n_{\text{moles}} N_A \left(\frac{1}{2} mv_{\text{total}}^2\right)_{\text{average}} \\ PV &= \frac{2}{3} n_{\text{moles}} N_A \left(KE_{\text{per atom}}\right)_{\text{average}} \end{aligned}$$

This is a remarkable relation, because we can compare it to the empirically determined ideal gas law ($PV = nRT$) and write:

$$RT = \frac{2}{3} N_A \left(KE_{\text{per atom}}\right)_{\text{average}}$$

or

$$\left(KE_{\text{per atom}}\right)_{\text{average}} = \frac{3}{2} \frac{R}{N_A} T$$

If you tell me the temperature, I can tell you the average kinetic energy per atom in a gas.

Turning this around, we can use it to tell us what temperature really is. Temperature is in fact a measure of the average energy per atom. It doesn't matter what kind of atoms you have, heavy or light, the temperature is a measure of the average kinetic energy of the particles which make up an object.

This equation also suggests that we might want to define a new constant, called the Boltzmann constant:

$$k_B = \frac{R}{N_A} = 1.38 \times 10^{-23} \frac{\text{Joules}}{\text{Kelvin}}$$

Using this constant, we can nicely write the average kinetic energy per atom in a gas as:

$$\left(KE_{\text{per atom}} \right)_{\text{average}} = \frac{3}{2} k_B T$$

Since typical temperatures are around 300 K, this tells you that the typical kinetic energy of an atom in a gas is:

$$\left(KE_{\text{per atom}} \right)_{\text{average}} = \frac{3}{2} \left(1.38 \times 10^{-23} \frac{\text{Joules}}{\text{Kelvin}} \right) 300 \text{ Kelvin} = 6.2 \times 10^{-21} \text{ Joules}$$

This seems like a tiny energy. But remember, atoms have very, very low masses. Let's imagine that this is a Helium atom. What would its velocity be? The mass of a Helium atom is about 4 amu, or 6.6×10^{-27} kg, so:

$$\begin{aligned} \frac{1}{2} m_{\text{He}} v_{\text{He}}^2 &= KE \\ v_{\text{He}} &= \sqrt{\frac{2KE}{m_{\text{He}}}} = 1366 \frac{\text{m}}{\text{s}} \end{aligned}$$

So at room temperature, each atom inside a Helium balloon is zipping along at about 1366 m/s. In more familiar units, that's a bit more than 3000 mph, so they're pretty peppy. If the gas is made of heavier atoms (Argon say, or N_2 molecules), the velocities are lower. But they're always quite high compared to the macroscopic velocities we're used to.

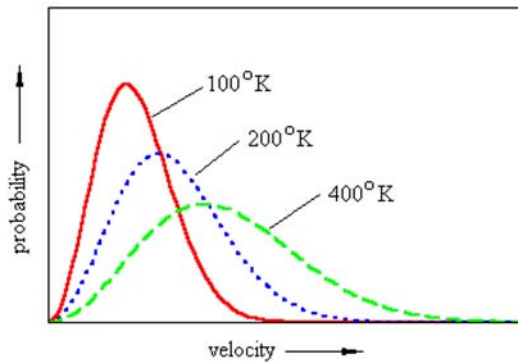
What is temperature really?

This "kinetic theory" discussion of ideal gases provides a way to show that, for this case, temperature is a measure of the average kinetic energy of particles in a gas. This is a very simple case. In particular, for an ideal gas like this, kinetic energy is the only form of energy the atoms can have. In more complex cases, like a solid, atoms can have both kinetic and potential energy. Inside a solid, atoms oscillate around their positions of equilibrium. During these oscillations they have both kinetic energy and potential energy. In these more complex cases, temperature is still a measure of the average energy of the atoms in a material, but now that energy appears in more forms than just kinetic energy.

Two objects which "have the same temperature" have the same average energy in each of their tiny parts.

This doesn't imply that all the atoms in the material have exactly the same energy. In fact the energy of each atom changes all the time as they bounce around against one another. What stays the same is the *average* energy of the atoms. At any given moment, the atoms in a gas will have a distribution of energies, some higher, some lower. Since there are typically so many atoms, this distribution is actually very stable.

The picture below right shows, schematically, what the distribution of particle velocities might look like. As you increase the temperature, the average speed increases, and the distribution of speeds becomes broader. As time goes on, atoms change speed, some going faster, some slower. But the distribution of speeds, and so its average, remain the same.



We have seen that temperature is a measure of the average amount of energy in each of the particles which makes up a material. Every atom in a material at temperature T will have an average kinetic energy of $\frac{3}{2}k_B T$. At room temperature this is about 6×10^{-21} J.

These thermal motions are a fundamental fact about matter. Atoms are *always* in motion. In gases they fly about freely, most of the time completely unconnected to one another. In liquids they slip around over one another, exchanging positions all the time but not completely escaping. In solids they remain mostly locked in place, but oscillate violently around their equilibrium positions.

Thermal energy, and this ubiquitous unending motion, is incredibly important for life. Life is based on a continuous flow of energy. Living things, yourself included, exist by continuously taking in energy, changing its form, and sending it back out again. Much of what you take in is converted to thermal energy, and many of the mechanisms in your body rely heavily on this simple thermal motion to do what they do. Before we explore the details of this, we need to address a deep and fundamental question about thermal motion, conservation laws, and what happens in the world.

13.4 Why do things happen the way they do: particle velocities

If we examine the velocities of these particles in a gas, we find they fill out a very particular distribution which is very stable if there are even a modest number of atoms. This fact raises a really big question. What is it that “makes” these atoms take on this particular distribution of velocities? Why does this particular distribution always emerge?

The conservation of energy and momentum *don't* require this distribution. They only insist that the total momentum and energy remains fixed. The conservation laws alone would be just as happy with all the energy and momentum in one atom, with all the others at rest. Conservation laws are not the influence which determines this distribution of velocities. What does?

This question is actually extremely broad. There are many things which *could* happen in physics, but which never do. These are processes allowed by the conservation of energy and momentum, but which never occur. Sliding a book to a stop across a table is an example.

When you do this, the kinetic energy of the book is converted to thermal energy in the table and the book. Conservation laws would *allow* this thermal energy to come back together and reenter the book. If this happened, you might set your book on the table, have it pick up thermal energy from its surroundings and suddenly start sliding away. That process is allowed, but it never happens. Sliding the book to a halt is something we'd call an "irreversible" process.

Irreversibility and chance

Energy flow constraints allow many transformations which never occur. How do we determine which ones will and which won't? Doing this is the subject of statistical mechanics, an enormously important branch of physics which answers just this question. How do you figure out not only what *can* happen, but what *will* happen?

Statistical physics relies on one fundamental assumption:

All possible outcomes are equally likely

To see how this very reasonable sounding assumption might be consistent with the irreversibility we just discussed, in which some outcomes never happen while others always do, we'll consider a simple model system. Imagine a cube filled with some ideal gas atoms rattling around in it. This is a closed box. Neither atoms nor energy ever enters or leaves it.

We begin by defining microstates and macrostates for the system. The "microstate" is simple. It's just the precise condition of all the atoms in the box: a complete detailed description at some instant. The idea of a macrostate is a little more subtle. A "macrostate" is some collective feature of the distribution of positions or velocities of the particles. Examples include:

- Average particle kinetic energy (temperature)
- Average pressure on the wall
- Average position of the atoms
- What fraction of the atoms are on the top half of the box

In general, macrostates like this are the sort of thing we might actually observe about such a box. There's no way to measure exactly where each atom in a gas is, along with exactly how each is moving. But it is perfectly possible to measure the temperature, determine approximately how much gas is on the top of the box, etc.

Next we need to think a bit about timescales. We want to see what will happen if we let this system evolve, so we have to wait long enough to let the system rearrange itself. Each time we check to see what the system has done (where the atoms are) we should give it time to rearrange itself before we check again.

How long do we wait? The typical time for rearrangement is comparable to the time it takes an atom to cross the box: L/v_{av} . We've seen that atoms travel quite fast in a gas, so we won't have to wait long. If our box is 10 cm across, we might need to wait $0.1 \text{ m} / 1000 \text{ m/s} = 10^{-4} \text{ s}$ or so. The only reason to raise this point is that some systems, big ones like galaxies, take *much* longer to rearrange themselves. What happens in these large slow systems is qualitatively the same, but the timescales are a lot longer.

Now we start looking at the box over and over again, each time waiting long enough for the atoms to rearrange themselves. If we do this, every possible arrangement of particles, every microstate, will eventually occur. In fact, they're all exactly equally likely. What about the macrostates? How often, for example, will we find all the particles in the top half of the box?

To figure this out, we need to find out what fraction of all the equally likely microstates corresponds to this macrostate. The probability that this macrostate will be observed is then equal to

$$P_{top} = \frac{(\text{number of microstates for this macrostate})}{(\text{total number of microstates})}$$

Let's work this out for our simple "all particles on the top of the box" macrostate. The answer depends on the number of particles. Consider one atom first. This atom can be on top or on the bottom. One of these two possible microstates corresponds to all atoms on top, so:

$$P_{top}(1 \text{ atom}) = 0.5$$

If there are two atoms, it becomes more complicated. Now you could have

- Both on top (1 arrangement)
- One on top, one on the bottom (2 arrangements, one with each on top)
- Both on the bottom (1 arrangement)

So now the probability of having them all on top is:

$$P_{top}(2 \text{ atoms}) = \frac{1 \text{ microstate}}{4 \text{ total microstates}} = 0.25$$

You should do the same for 3 and four atoms in the box. Label them A,B,C, etc. and figure out how many total arrangements there are for three and four in the box. You should find that:

$$P_{top}(3 \text{ atoms}) = 0.125 \quad \text{and} \quad P_{top}(4 \text{ atoms}) = 0.0625$$

In fact, there is a general rule for this:

$$P_{top}(N \text{ atoms}) = \left(\frac{1}{2}\right)^N$$

So this particular macrostate, with all the atoms on the top, becomes increasingly unlikely as the number of atoms increases. The microstate which makes this macrostate is not *intrinsically* less likely than any

other microstate. In fact it's exactly as likely as any other microstate. But as the number of atoms increases there are more and more different microstates, only one of which corresponds to this macrostate. This particular macrostate gets very unlikely as the number of atoms increases.

There is another macrostate, however, which becomes increasingly likely. Think back to the 4 atom case. How often do you have half the atoms on top and half on the bottom? There are 16 total microstates for these four distinguishable atoms. Six of these correspond to half on top and half below. So the probability of finding the half-and-half macrostate is $6/16 = 0.375$. It's much more likely that the all-on-top macrostate.

This is the key idea in statistical mechanics. All microstates are equally probable. Every allowed outcome is equally likely. But some *macrostates* correspond to many microstates, while others correspond to very few. Random wandering among the possible microstates then guarantees that some macrostates will happen more often than others. In this sense, having all the kinetic energy in a sliding book is a macrostate which involves few microstates, while having that energy spread out as heat in the table involves many microstates. Once you move from the rare, all concentrated in the book state, into one of the incredibly many ways you can have the energy spread out, you never wander back into the rare state of having all the energy in the book.

How can this be irreversibility?

Nothing we have said would seem to explain irreversibility. According to this, all we have to do to see the book take off across the table is wait long enough. In our four atom example, we'd only have to wait about 16 rearrangement times to see all the atoms on top. How long might this be? Imagine our atoms rearrange themselves a million times a second. For our 4 atom system, we would, on average, see them all on top every 16 microseconds. This would happen about 62,000 times a second. It's pretty common.

What if we increase the number of atoms to a reasonable number? If you have even 80 atoms, still a very tiny number of atoms, the probability of finding them all on one side is $1/1200000000000000000000000$, or about 8.7×10^{-25} . These atoms, zooming along, rearranging themselves a million times a second, will all be on one side once every 10^{18} seconds or so. That's about 30 billion years...Unfortunately the universe is only about 13.7 billion years old so far, so it probably hasn't happened yet.

To be more realistic, 1 cm^3 of air contains about 2×10^{19} atoms. How often will you find *them* all on one side of the box? Never. Not because you can't. Just because the odds are colossally against it.

So the following simple process, push all the atoms in a 1 cm^3 box to one side, then let them go, is, despite the equal probability of every microstate, completely irreversible. The sliding of a book to a halt, while more difficult to calculate, is just the same. There's only one way to have all that energy in the kinetic energy of the book, with all its atoms sliding along together. But there are quadzillions (that's the technical term) of ways to spread that energy out in random motion of atoms in the book and the table. So sliding to a halt is irreversible. Not because it can't reverse, but simply because it won't.

Using this in statistical physics

This toy example gives us a good idea of how statistical physics works. To predict what will happen you define macrostates you want to explore, then figure out how to count the allowed microstates which

correspond to each macrostate. Comparing then tells you how likely each macrostate is. Those macrostates which are made from many microstates are the most likely. Usually the tricky part is counting the microstates.

In general what you find is that any state which spreads the energy out equally among all the parts is vastly more likely than a state with all the energy concentrated in a subset of the matter. This idea, that energy will spread out as much as possible, is called the principle of equipartition. It isn't meant to suggest that energy is spread perfectly equally, but that it is spread out in some calculable distribution (some atoms have more energy, some less, and the distribution is stable).

To summarize: Random chance is what decides which among the allowed possible outcomes will actually occur. How long will it take the final outcome to emerge? This depends on how rapidly the system rearranges itself. The progress from order (like all on one side) to disorder (like evenly spread) happens by chance, not by force or design. In any system you can measure, it is so likely as to be practically inevitable. Once you get into this kind of disordered state, you won't leave it. It is a stable equilibrium state.

There are cases, lots of them, which involve change from disorder to order. But in every case, this happens because the system is open, and able to exchange matter and energy with the rest of the universe. If you bring energy into a system, you can cause it to move into a more ordered state. For example, we could push that gas into the upper half of our 1 cm³ block. But if you leave it alone, if the system is truly isolated, it will inevitably progress toward disorder.

Entropy and statistics

You have probably heard about “entropy”, and that entropy always increases. There is a quantitative connection between the ideas here and entropy. In fact, the clearest definition of the entropy of a system in a particular macrostate is:

$$S = k_B \times \ln(\# \text{ of microstates equivalent to this macrostate})$$

So when people say entropy increases, they just mean that the system will, by chance, migrate toward macrostates represented by many microstates. Typically, high entropy macrostates are those in which energy in the system is spread out as much as possible. For an isolated, homogeneous object, this means the maximum entropy will be reached when energy is spread uniformly through it; when all parts of it reach the same temperature. Once this happens, the system reaches its maximum entropy state and stops changing.

A Quick Summary of Some Important Relations

Ideal gases as a model material:

Ideal gases exert pressure on their surroundings. Observations show that the pressure, volume, and temperature of a number n moles of gas are related by the ideal gas law:

$$PV = nRT$$

In this relation, R is the ‘universal gas constant’, which has a value of 8.3 J/mole*K.

Temperature and the kinetic energy of atoms:

Combining the ideal gas law with an atomic picture of a gas allowed us to define temperature in terms of atomic kinetic energy:

$$KE_{\text{atom}}^{\text{average}} = \frac{3}{2} k_B T$$

When we say an object has a certain temperature, we are actually specifying the average kinetic energy of each of its atoms. Gases taught us this, but it is true for solids and liquids as well.

Velocities change, the distribution of velocities does not:

In a gas, an enormous number of atomic collisions occur each second, each trading momentum and energy from one atom to the next. These change the velocities of each, but so long as there are many atoms, the distribution of velocities will be very stable indeed.

The fundamental principle of statistical physics:

To understand why many outcomes allowed by conservation laws never happen, we use the fundamental principle of statistical physics: that all allowed microstates are equally probable. When some macrostates correspond to many more microstates than others, those macrostates will become extremely likely.

Using this principle to make predictions requires that you can count how many allowed microstates correspond to each macrostate. Doing this can be complex, but is possible, and allows us to precisely predict, for example, the distribution of velocities seen for the atoms in a gas.

Entropy and the number of microstates:

The entropy of a system is determined by the definition:

$$S = k_B \ln(\Omega) \quad \text{where } \Omega = \text{number of equally likely microstates in this macrostate}$$

Since systems always end up in the macrostates with the most microstates, entropy always increases.

14. Using chaos: diffusion and life

- 1) Diffusion: moving things with random motion
 - i. Particle size and mean free path
 - ii. The diffusion coefficient and time scales
- 2) As easy as breathing: getting oxygen into your cells
- 3) An alternative view of diffusion
- 4) Diffusion across membranes and osmosis
 - i. Consequences of osmosis
 - ii. Osmosis in culture

Physics for the Life Sciences: Chapter 14

14.1 Diffusion: moving things with random motion

We have seen in the last chapter that completely random processes, like atoms rattling around in a cube, can lead to inevitable conclusions. These things happen by chance, not by force or design, and yet remain inevitable. As it happens, life uses events which occur through random processes very extensively. This is especially important and apparent in two related processes, diffusion and osmosis. Later we'll discuss more generally how this harnessing of random motion, of what's going to happen no matter what, is behind all of life.

We have seen how random motion causes atoms released in just one region of a box to inevitably spread uniformly throughout it. Life takes advantage of the spreading caused by random motion very extensively to move things around. The process is called "diffusion", and it relies on the random thermal motions of particles. When particles are at a temperature T they have a typical velocity (taken here as the root-mean-square velocity) which depends on the temperature. We find this by recalling the relation between temperature and kinetic energy:

$$\frac{1}{2} m_{\text{atom}} v_{\text{atom}}^2 = \frac{3}{2} k_B T$$

This can be written in another form, expressed in terms of the molar mass:

$$v_{\text{RMS}} = \sqrt{v^2} = \sqrt{\frac{3k_B T}{m_{\text{atom}}}}$$
$$m_{\text{atom}} = \frac{M_{\text{mole}}}{N_A} \quad \text{and} \quad k_B = \frac{R}{N_A}$$
$$v_{\text{RMS}} = \sqrt{\frac{3RT}{M_{\text{mole}}}}$$

For any material, this v_{rms} varies like the square root of the temperature divided by the molar mass. To double v_{rms} by heating things up you have to quadruple the temperature. To reduce v_{rms} by half by changing the substance you would need to increase the molar mass by a factor of four. Most of life exists

in a very narrow temperature range, something like $273\text{ K} < T < 373\text{ K}$, so the typical v_{rms} doesn't vary much because of temperature changes. But the molar mass of substances you might find in living things does change a lot; by factors of about 20,000. This has a strong effect on their typical thermal velocities. Here is a table showing this large variation, from 500 m/s for a small molecule like N_2 to 2.6 cm/s for a virus.

Table 2.1. Root Mean Square Velocities for Various Molecules at $t = 15^\circ\text{C}$ ($T = 288^\circ\text{K}$).

Molecule	m (g)	M (g/mol)	v_{rms} (m/s)
Hydrogen H_2	3.35×10^{-24}	2.016	1880
Helium He	6.65×10^{-24}	4.002	1340
Nitrogen N_2	4.65×10^{-23}	28.02	506
Oxygen O_2	5.32×10^{-23}	32.0	474
Mercury Hg	3.33×10^{-22}	200.6	186
Macromolecules	1.67×10^{-20}	10^4	26 m/s
	1.67×10^{-18}	10^6	2.6 m/s
Viruses	1.67×10^{-16}	10^8	26 cm/s
	1.67×10^{-14}	10^{10}	2.6 cm/s

Particle size and mean free path (MFP)

To figure out how quickly random motion will get things from one place to another, we need to know not only how fast things move, but also how far they go between collisions. If they move fast, but are continually running into other things, they won't get far. On the other hand, if they *never* hit anything else, they will zoom along at constant speed. Now a truly ideal gas is like that. The atoms in such a gas are supposed to never interact. But real atoms do. We can work out how far they go between collisions from the ideal gas law:

$$V = \frac{nRT}{P} \quad \text{or} \quad V_{\text{mole}} = \frac{RT}{P}$$

$$\frac{V}{\text{Particle}} = \frac{V_{\text{mole}}}{N_A} = \frac{RT}{N_A P}$$

$$V_{\text{swept out}} = \pi(2r)^2 d_{\text{MFP}} = \frac{V}{\text{Particle}}$$

$$d_{\text{MFP}} = \frac{RT}{N_A P \pi (2r)^2}$$

$$d_{\text{MFP}} = \frac{8.3 \frac{\text{J}}{\text{mole} \cdot \text{K}} \times 300\text{K}}{6 \times 10^{23} \times 10^5 \frac{\text{N}}{\text{m}^2} \times \pi \times (2 \times 10^{-10} \text{m}^2)^2}$$

$$d_{\text{MFP}} \approx 3 \times 10^{-7} \text{m}$$

Basically, this derivation finds the volume per particle, then says this volume is equal to the volume “swept out” by the particle as it travels from one collision to the next. This provides an estimate of the distance between collisions, what's called the “mean free path” of the particles. You can see that it

depends on temperature, pressure, and the radius of the atom r . For typical values in a cold gas, this mean free path is around 10^{-7} m; not very far at all. You get the typical time between collisions from

$$t_{\text{mean}} = \frac{d_{\text{mfp}}}{v_{\text{rms}}}$$

and putting in typical numbers you find times around 10^{-10} s.

So when you picture atoms rattling around in a gas you should imagine they have around 10 billion collisions a second, each time traveling around $0.1 \mu\text{m}$ between collisions. Atoms moving like this conduct what's called a "random walk", something like what's illustrated in this picture. During this walk, they gradually move away from where they started. Since it is a random process, you can't predict *exactly* how far or in which direction any particle will move, but you can statistically predict how far particles will move on average.



The diffusion coefficient and timescales

It turns out this distance is characterized by the "diffusion coefficient", which for an ideal gas is sometimes defined as:

$$D = \frac{1}{2} v_{\text{rms}} d_{\text{mfp}}$$

Notice that it depends both on how fast things are moving and how far they go between collisions. In fact the diffusion coefficient for different situations is often measured, rather than calculated from first principles. The units of the diffusion coefficient can be seen from the definition to be m^2/s .

Using this definition for D , the rate at which particles move away from where they start due to this random walk is given by:

$$r_{\text{rms}} = \sqrt{6Dt}$$

$$x_{\text{rms}} = \sqrt{2Dt}$$

In these relations, r_{rms} is the root-mean-square distance away from the starting point in three dimensions, while x_{rms} is the root-mean-square distance in a single direction (like x for example).

What's important to notice about this? First, the distance traveled varies like the square root of the time. So if you want a particle to diffuse twice as far it will take four times as long. Second, the distance depends on this diffusion coefficient, and hence on v_{rms} and d_{mfp} . As a result, small, light atoms will

diffuse faster than large atoms. In addition, diffusion will happen faster when the temperature is larger or the density is lower.

Here are some tables of diffusion coefficients:

Molecule	Temp (C)	D (m ² /s)
H ₂	0	6.4x10 ⁻⁵
H ₂ O (Vapor)	8	2.4x10 ⁻⁵
O ₂	0	1.8x10 ⁻⁵
CO ₂	0	1.4x10 ⁻⁵

Diffusion in air

Molecule	Temp (C)	D (m ² /s)
O ₂	20	1.0x10 ⁻⁹
Glucose	20	6.7x10 ⁻¹⁰
Hemoglobin	20	6.9x10 ⁻¹¹
DNA	20	1.3x10 ⁻¹²

Diffusion in water

Given these numbers, you might expect a molecule of O₂ to diffuse about 1 cm in a second in air, while it would diffuse through only 70 μm during the same second in water.

So diffusion is a process which can move atoms and molecules from one place to another just by allowing them to rattle around. No mechanism is needed to make it happen. This mechanism is used very extensively by life, but it is limited to working over relatively short distances.

One particularly important example is O₂, which is used in your body for metabolism. O₂ can be delivered by diffusion, but it diffused rather slowly through water. For example, in water at body temperature, it takes:

- 2x10⁻⁴ s to go 1 μm
- 170 s to go 1 mm
- 4.6 hours to go 1 cm
- 5.4 years to go 1 meter

Since your cells can't wait five years for oxygen to travel from your lungs to your legs, some method of transport other than diffusion is required, and that's why you have this complex circulatory system which moves bulk amounts of oxygen loaded blood around.

To have diffusion be effective, you need to keep the distance scales short. Diffusion is the primary method of transport within a cell. Inside the cell, you get key molecules from one place to another just by letting them go. They diffuse around, ramming into things randomly, until they get where they're supposed to go.

The fact that diffusion is so important in cellular transport is the reason cells are so small. If they were larger, say 1 cm in size, it would take much too long for important materials to diffuse from one place to another. If an organism wants to grow bigger, it has to develop methods of transport which carry things

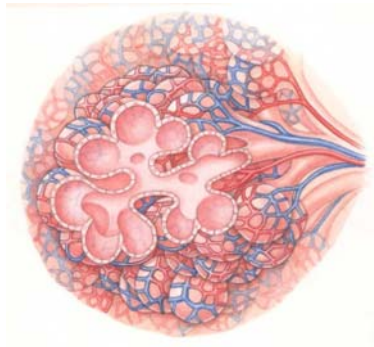
long distances much more effectively than diffusion. Most cells are on the 1-10 μm size scale so that they can efficiently use diffusion as a mechanism of transport.

14.2 As easy as breathing: getting oxygen to your cells

The transport of oxygen from the atmosphere into an organism, then through it to the cells where it must be used, is a great example of how different approaches to a fundamental physical constraint have been explored by evolution. Let's consider the steps.

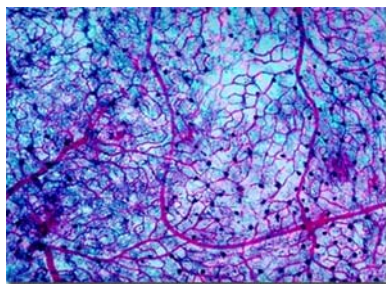
1: First, to get oxygen from the atmosphere into your blood stream it must diffuse across a membrane. That's easy, just make a thin membrane and let it go. There is a problem though, at least for us large organisms. Remember that surface area is proportional to size^2 , while volume is proportional to size^3 . The amount of O_2 you can get per unit time depends on surface area, while the amount you *need* depends on volume. So as organisms get larger, they need to somehow manufacture more and more oxygen diffusing surface area.

There are at least three general solutions. One is to stay small. This works for insects and other smaller organisms. They mostly don't have to breathe. They can just wait around for O_2 to diffuse in. A second approach is gills. These big feathery structures have enormous amounts of surface area given their volume. Hanging them out in the water is enough to allow lots of O_2 to diffuse in. Our approach, along with most land animals, is to make internal lungs which are packed full of tiny sacs called alveoli. The reason for the sacks is to maximize the surface area which is available for O_2 diffusion. These little sacks create enormous amounts of area and allow red blood cells to get really close to the surface so that oxygen can get to them.



2: Once you get the oxygen into a carrier (like a red blood cell) you need a very efficient transport system for moving it to where it's needed. This is what circulatory systems are for. If you didn't need one, you certainly wouldn't have one, because they fail so easily. Your arteries bring together a huge number of capillaries to allow efficient pumping off the remote parts of the body.

3: Once there, you again have to put the oxygen carrying red blood cells close enough to the target cells to allow diffusion to deliver the oxygen. That's what the incredibly fine capillary system through the rest of your body is for.

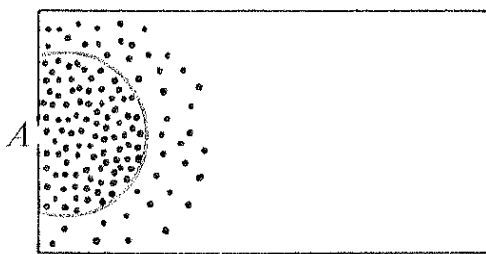


14.3 An alternative view of diffusion

In the discussion above, we considered how diffusion, the random motion associated with thermal energy, can cause an atom to wander away from its starting point. This wandering can be used for transport, but to be *useful* it must happen where there is a density gradient.

To see why, imagine a box filled with N_2 molecules. As time goes by, each molecule will wander away from its original position according to the equations described above. But since every molecule is wandering, the average density of molecules at any point will remain exactly the same. No *net* transport will take place. This is just another way of saying that this uniform distribution is an equilibrium condition; the average properties, the macrostate of this, will not change.

To have some real transport, some flow of molecules, which is driven by diffusion, you need to have density gradients. The simplest case is the old cube of gas. If you start all the atoms on one side so the density on one side is high and on the other it is zero, you will get diffusion driven net flow until the density is uniform.



More generally you might have a case with a smooth density variation like this. In this case you would measure the density gradient by seeing how the density changes with position at each point. In the general case you would write the density as a function of position $\Phi(x, y, z)$ and then measure how rapidly it changes in each direction. In this case you would describe the density gradient measuring the derivative of the density along each direction using what are called “partial derivatives”. You will learn how to do this when you learn vector calculus.

To see basically how this works, let’s consider a one dimensional version. Imagine a case of a tube along which the density of a substance varies as $\Phi(x)$ and remains fixed in time. This is what happens in steady state diffusive flow. In this case we expect the net flow of material to be described by Fick’s Law:

$$J = -D \frac{d\Phi(x)}{dx}$$

In this equation D is the diffusion coefficient we encountered above, $d\Phi/dx$ is the density gradient (the rate of change of density with position), and J is the flow rate or current of particles. J is measured in units of number / area*second.

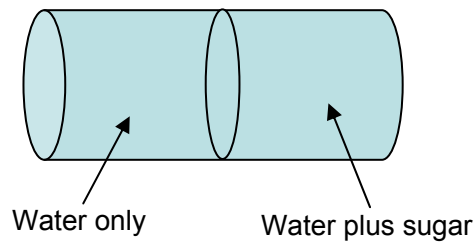
What does Fick's Law tell us? First, the rate at which particles flow due to diffusion depends on the diffusion coefficient D . Recall that D depends on temperature and the properties of what's diffusing through what material. Second, Fick's Law tells us that there will be flow only where there is a density gradient, and the net flow will be largest where the gradient is largest. Notice that this depends on a gradient of density in just the same way that heat flow depended on the gradient of the temperature. In fact, Fick's law is identical to the equation for thermal conduction. Heat flow is mathematically identical to diffusion.

14.4 Diffusion across membranes and osmosis

A key task for life is to separate itself from its environment. By doing this, life can control its conditions, ensuring that sudden changes in the outside world do not instantly affect it. Life does this isolation with membranes. There are many kinds of biological membranes. Some surround cells, isolating them to maintain a controlled environment inside. Others define parts of the cell, like peroxisomes, the nucleus, and the Golgi apparatus. Most are constructed of lipid bilayers; back-to-back sheets of fatty molecules. Typical thicknesses range from 0.5-4.0 nm, so they are very thin indeed.

The main purpose of membranes is to control the environment within by selectively allowing material to cross. Biological membranes are selectively permeable; they allow some substances to pass through them while blocking others. They are often called "semipermeable" membranes, though this name is perhaps misleading. This property, when combined with diffusion, allows osmosis, and it has a number of interesting properties.

To get a sense of this, consider a toy case in which such a semipermeable membrane separates two compartments in a cylinder. On the right, the pressure is created by both water and sugar molecules striking the walls. On the left the pressure is created only by water molecules hitting the walls. The pressure starts out equal, so there is a higher density of water on the left hand side. There must be, to make up for the pressure made by the sugar molecules on the right. Now imagine what happens if the membrane in the middle is semipermeable, and lets through water but not sugar.



At the start, there will be a density gradient of water across the membrane. This density gradient will allow diffusion to transport water from the left to the right, until the density of *water* will be equal on both sides. At this point the pressure created by water molecules will be the same on both sides; they will strike the membrane equally from left and right. But the *total* pressure on the two sides will now be *different*. On the left you have both water and sugar molecules hitting the walls, while on the right you have only an equal number of water molecules hitting the walls. The pressure on the right will be higher by the amount contributed by the sugar molecules.

This final state is described by writing:

$$P_{\text{inside}} = P_{\text{outside}} + P_{\text{osmotic}}$$

This P_{osmotic} is just the extra pressure created by the material which can't diffuse across the membrane (the sugar in this case), and the symbol π is commonly used for P_{osmotic} .

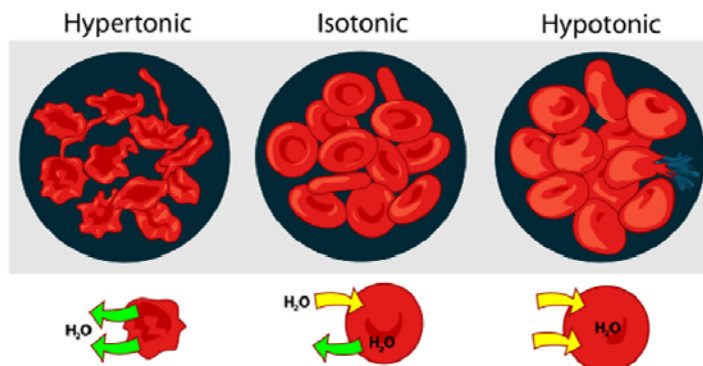
Can we determine how large this extra pressure is? Doing this precisely is tricky, but as long as the solution is dilute (there is little sugar in the water), we can actually use the ideal gas approximation and come pretty close:

$$P_{\text{sugar}} = \pi = \left(\frac{nRT}{V} \right) = \left(\frac{n}{V} \right) RT = c_{\text{molar}} RT$$

Where c_{molar} is the molar concentration, the number of moles per unit volume. From this we see that the osmotic pressure increases directly with the solute concentration c_{molar} , and also with temperature T .

Consequences of osmosis

Many membranes are permeable to water but not to larger molecules. When this happens, water diffuses across the membranes selectively as we have seen above. Sometimes this happens in “hypotonic” (too much tension) environments, where the concentration of the solute is larger inside the membrane. In these cases, osmotic pressure builds inside the cell. This can, in extreme cases, lead to membrane rupture and cell death. In “hypertonic” (too little tension) environments, the opposite is true, and water diffuses out of the cell. When the solute concentration is balanced, conditions are “isotonic” and the flow of water into and out of the cell is balanced.



Osmosis in culture

Interestingly, the word osmosis has been picked up as a popular term. Here's the definition of the popular version from the OED: "A process resembling osmosis, *esp.* the gradual and often unconscious assimilation or transfer of ideas, knowledge, influences, etc." Things that are adopted "by osmosis" somehow happen by themselves, without being forced or planned. This is a remarkably apt way of remembering what's happening in diffusion and osmosis. These aren't processes driven by pumps or will, they occur as the inevitable outcome of a lot of random chance.

The association of osmosis with some kind of mysterious influence goes back to its discovery in the 19th century. At this time, the atomic nature of matter was not completely accepted, and certainly not understood, so osmosis seemed almost like magic. You can get the sense in this sentence, from British scientist Thomas Graham in 1854: "This flow of water through the membrane I shall speak of as osmose, and the unknown power producing it as the osmotic force". This *unknown power* is the idea that caught the popular mind and led to adoption of this phrase.

A Quick Summary of Some Important Relations

Random thermal motion:

Atoms in all matter are constantly moving and colliding with one another. They have typical velocities which can be characterized by a root-mean-square value, and travel typical distances between collisions characterized by mean-free-path:

$$v_{\text{RMS}} = \sqrt{\frac{3k_B T}{m_{\text{atom}}}} = \sqrt{\frac{3RT}{M_{\text{molar}}}}$$
$$d_{\text{MFP}}^{\text{Ideal gas}} = \frac{RT}{4\pi N_A P r^2}$$

Random motion and transport through diffusion:

This motion and the collisions which it engenders leads to diffusive transport characterized by a diffusion coefficient:

$$D = \frac{1}{2} v_{\text{RMS}} d_{\text{MFP}}$$

This random motion leads to a wandering motion that moves a typical distance given by:

$$r_{\text{RMS}}^{3D} = \sqrt{6Dt} \quad x_{\text{RMS}}^{1D} = \sqrt{2Dt}$$

Density gradients and diffusive flow:

When the density of atoms is not the same everywhere, random motion leads to a net flow of atoms which can be characterized as a ‘current’. The number of particles per unit area passing through an area at a particular location is described by Fick’s law:

$$J(x) = -D \frac{d\Phi(x)}{dx}$$

Notice that the current at each point depends on the density gradient at that point: this is a local relation.

Osmosis and osmotic pressure:

Osmosis is diffusion across a semipermeable membrane. It can create a pressure difference across the membrane which can become important. For low concentrations, this can be estimated by:

$$P_{\text{osmotic}} = c_{\text{molar}} RT$$

15. Keeping cool and staying warm: thermal transport and life

- 1) Temperature and thermal expansion
 - i. Volume expansion
- 2) Heat capacity
 - i. Predicting molar heat capacities
 - ii. Diatomic gases
 - iii. Solids and more complex materials
 - iv. Molar heat capacity and specific heats
 - v. Heat capacity and energy storage
- 3) Changes of state and heat of transformation
- 4) Moving heat around
 - i. Conduction
 - ii. Convection
 - iii. Radiation
 - iv. The nature of thermal radiation
- 5) Size, thermoregulation, and life
 - i. Benefits and methods of thermoregulation
 - ii. Thermoregulation on land
 - iii. Thermoregulation in water

Physics for the Life Sciences: Chapter 15

In this chapter we'll expand our discussion of thermal energy and consider what happens when heat energy is added to liquids and solids. In these cases, where atoms are not freely flying but are attached together by interatomic forces, the introduction of heat is a little more complicated than it is with gases, though all the same main ideas apply. In this chapter, we will focus on how thermal energy moves around on a macroscopic scale, in objects of ordinary size, like multicellular organisms.

15.1 Temperature and thermal expansion

In ideal gases, temperature appears only as motion, and as we've seen the energy put into a gas as heat spreads smoothly throughout, giving rise to a distribution of velocities with the average kinetic energy of each atom of $\frac{3}{2}k_B T$. In a solid too, heat energy added will spread out among the atoms, but now they are not free to fly around. Instead they oscillate around their positions of equilibrium. In these oscillations, each atom trades energy between kinetic and potential energy, on average splitting that energy equally between the two. So in addition to kinetic energy, each atom has also potential energy associated with the stretching of the bonds which hold them together. This oscillation around equilibrium, which increases with temperature, has many consequences. The first is thermal expansion.

You can think of atoms in a solid as being attached by springs. If two such atoms are connected by a perfect Hooke's law spring, then the average distance between them will be the same whether they oscillate or not, because the oscillations will move them apart and together in exactly equal amounts. In fact though, the forces which hold atoms together are not perfect, linear, Hooke's law forces. Instead they are asymmetric. Usually, it is harder to push two atoms closer together by a distance Δx than it is to

stretch them apart by the same distance. It's a little easier to pull the atoms apart than it is to squash them together.

When atoms start oscillating in such a material, the average distance between them becomes larger. The solid expands. Exactly how much depends on just how asymmetric the “springs” connecting the are, which depends on the nature of the chemical bonds.

This **thermal expansion** of a solid rod with length L can usually be well described by a phenomenological relation:

$$\Delta L = \alpha L \Delta T$$

where α is the “coefficient of thermal expansion” for this material, and ΔL is the change in length of the material. Notice that the change in length ΔL depends on the length L . This is because *each* bond in the material becomes a little longer. So if there are more bonds, there is a larger increase in length. As a result, it's often more useful to consider the fractional change in length:

$$\frac{\Delta L}{L} = \alpha \Delta T$$

The table to the right lists these thermal expansion coefficients for a few materials. Notice that they are small. If you change the temperature by 1° C, then a stainless steel rod will undergo a fractional change in length:

$$\frac{\Delta L}{L} = \alpha \Delta T = 1.7 \times 10^{-5}$$

or about 0.0017%.

Another important thing to notice is that they vary a lot. Why is this important? If you build something made of tightly fit parts of glass and aluminum, then change the temperature, the aluminum will expand about three times as much as the glass. If you hadn't counted on this in your design, parts may come apart or break. The stresses associated with thermal expansion can be quite spectacular. You can see this by comparing the thermal expansion described here with another way of changing the length: stress.

Material	α in $10^{-6}/\text{K}$ at 20 °C
Mercury	60
Aluminum	23
Stainless Steel	17.3
Glass	8.5
Pyrex	3.3
Invar	1.2
Diamond	1

Recall the stress-strain relation: stress = Young's modulus * strain. Putting this into symbols:

$$\frac{F}{A} = Y \left(\frac{\Delta L}{L} \right)$$

Now imagine you have a bar of aluminum fit tightly into a bracket. If you heat the bar it will expand according to the thermal expansion law above. If the bracket does not expand, then it must squeeze the bar inward. How hard must it push to prevent the bar from expanding? The stress-strain relation tells us:

$$\frac{F}{A} = Y \left(\frac{\Delta L}{L} \right) = Y(\alpha \Delta T)$$

For aluminum, $\alpha = 23 \times 10^{-6} / \text{K}$, and $Y = 70 \times 10^9 \text{ N/m}^2$, so this relation becomes:

$$\frac{F}{A} = \left(7 \times 10^{10} \frac{\text{N}}{\text{m}^2} \right) (23 \times 10^{-6} \text{ K}^{-1}) \Delta T = \left(1.6 \times 10^6 \frac{\text{N}}{\text{m}^2} \right) \Delta T$$

The bracket has to apply a pressure of $1.6 \times 10^6 \text{ N/m}^2$ to keep the aluminum from expanding if its temperature rises by just one degree! Because of this, differential thermal expansion is a major consideration for engineering in any case where an object has to survive significant changes in temperature.

One last thing to know; thermal expansion coefficients vary with temperature, often rather strongly, so as an engineer tries to deal with this issue they have to be careful to account for this as well.

Volume expansion

Thermal expansion doesn't just happen in one direction, instead materials expand in every direction. This makes their volume change. To track this, we'd like to know how the volume changes with temperature. This can be captured in a volume expansion coefficient so that:

$$\Delta V = \beta V \Delta T$$

How does this volume coefficient β relate to the linear expansion coefficient α ? To see, just calculate how the volume of a cube changes with temperature. Since we expect changes in length ΔL to be small, we can drop terms in ΔL^2 and ΔL^3 along the way.

$$\begin{aligned} V_{old} &= L^3 & \Delta V &= V_{new} - V_{old} \approx 3L^2 \Delta L \\ V_{new} &= (L + \Delta L)^3 & \Delta V &\approx 3L^2 \Delta L = 3L^2 \alpha L \Delta T = 3\alpha V \Delta T \\ V_{new} &= L^3 + 3L^2 \Delta L + 3L \Delta L^2 + \Delta L^3 & \Delta V &= \beta V \Delta T \\ V_{new} &\approx L^3 + 3L^2 \Delta L & \beta &= 3\alpha \end{aligned}$$

So the volume expansion coefficient for most solids is just about three times the linear expansion coefficient.

15.2 Heat capacity

We know the temperature of a material is related to the kinetic energy shared amongst its atoms. How much energy must we add to a material in order to change its temperature? The answer lies in our already established understanding of statistical physics.

When we dump energy into a material what will happen to it? Any energy we add will eventually get shared out equally among all the different locations and forms it can take. Each atom in the material will take some of the energy. Meanwhile, temperature is a measure of the *average* energy in each atom of the material. Since each atom takes energy, we to raise the temperature of the material we must raise the average energy of all the atoms; raising the temperature of a material with more atoms will require more energy.

We express this cost (how much energy is needed to raise the temperature) by defining the heat capacity of a material. This is defined in two ways, the “molar heat capacity” and the “specific heat capacity”. They’re simply related, but you must be careful to keep the two straight.

The molar heat capacity is defined by measuring how much heat you must add (ΔQ) to a given number of moles n_{moles} if you want to raise the temperature by an amount ΔT :

$$\Delta Q = n_{\text{moles}} C_{\text{molar}} \Delta T \quad \text{or} \quad C_{\text{molar}} = \frac{\Delta Q}{n_{\text{moles}} \Delta T}$$

The specific heat capacity is defined by measuring how much heat you must add (ΔQ) to a given mass of material m if you want to raise the temperature by an amount ΔT :

$$\Delta Q = mc_{\text{specific}} \Delta T \quad \text{or} \quad c_{\text{specific}} = \frac{\Delta Q}{m \Delta T}$$

The two are related in a simple way, because a mass m includes a number of moles $n = m/M_{\text{molar}}$. This implies:

$$c_{\text{specific}} = \frac{C_{\text{molar}}}{M_{\text{molar}}}$$

In general, the molar heat capacity is a more fundamental quantity. It essentially measures the amount of heat you must add *per atom* to raise the temperature. The specific heat involves both the atom-by-atom absorption of heat and the number of atoms per unit mass the material contains. As a result, molar heat capacities show some regularities which are hidden in specific heats.

Predicting molar heat capacities

If we understand what happens when we put energy into a system, we should be able to predict molar specific heats. If we put energy into an ideal gas, the only form it can take is kinetic energy. But if we put energy into a solid, it can appear as both kinetic and potential energy. So to make the average *kinetic* energy of atoms in a solid increase, we’re going to have to add more energy per atom than we would in

the gas; some of it will go into potential energy too. As a result, we expect molar heat capacity of solids to be higher. How much higher?

The available forms of energy internal to a material are sometimes called “degrees of freedom”, and at a given temperature each atom in a material will have an average energy of $\frac{1}{2}k_B T$ in each degree of freedom. For an ideal gas, these degrees of freedom are motion in the three directions, x, y, and z. Each of these takes up $\frac{1}{2}k_B T$, so the total energy per atom is $\frac{3}{2}k_B T$, as we have seen before. What does this make the molar heat capacity?

To raise the temperature by an amount ΔT , we have to give each atom $\frac{3}{2}k_B T$ of energy. If we do this for a mole of gas, we have to put in a total energy:

$$\Delta Q = N_A \left(\frac{3}{2} k_B \Delta T \right) \quad \text{or} \quad \frac{3}{2} N_A k_B = \frac{\Delta Q}{\Delta T}$$

Remembering that $N_A k_B = R$, the universal gas constant, we see that the molar heat capacity of an ideal gas should be $\frac{3}{2}R$, or about 12.5 Joules/mole \times Kelvin. For the monatomic noble gases (He, Ne, Ar, etc.) this is just what you find. From this you can see that the molar heat capacity for a material will in general be $\frac{1}{2}R$ for each degree of freedom in the system.

Diatomic gases

What about a diatomic gas like N_2 or H_2 ? A molecule like this is like a dumbbell, with the two atoms attached by a stiff spring. Here you can still put energy into translation, so you get $\frac{3}{2}R$ from that. There are other ways to store energy though. The dumbbell can rotate. Imagine it's aligned along the x-axis. For this, you can put energy in by rotating around either the y or z axis. That's two more degrees of freedom. You don't get energy into rotation around the x-axis because this diatomic molecule (like a little rod) is aligned along that direction, and rotations around this would hold little energy.

In addition, the two atoms can vibrate along the axis of the molecule (it gets longer and shorter). Because this motion involves oscillations around equilibrium, it includes both kinetic and potential energy, so that's two more degrees of freedom. As a result, there should be seven total degrees of freedom (3 for translation, 2 for rotation, and 2 for vibration), and the molar heat capacity should be $\frac{7}{2}R$.

At high temperatures, this is exactly right. All diatomic gases have $C_{\text{molar}} = \frac{7}{2}R$ when the temperature is high enough. But as you lower the temperature, the molar heat capacity drops; first to $\frac{5}{2}R$, then to $\frac{3}{2}R$. When this was discovered, it was a real mystery.

It appeared that, when the temperature was very low and the thermal energy small, it was impossible to put energy into the form of rotation or vibration, though it could still go into translation. So at low temperature this looked like the diatomic gas could only translate, just like a monatomic gas. Then as you raised the temperature, it would gradually become possible for the molecule to rotate, adding two more

degrees of freedom and increase the molar heat capacity to $\frac{5}{2}R$. Finally, it became possible for the molecule to vibrate, and the molar heat capacity rose again to $\frac{7}{2}R$.

This strange behavior was one of the earliest recognized signatures of quantum mechanics. The energies for rotation and vibration of these molecules are “quantized”. They can’t have just any values (from zero up) but can only take particular, well separated values. As a result, it is impossible for the molecule to absorb thermal energy into vibration until the typical amount of thermal energy around ($\sim k_B T$) is comparable to the jump from one quantum state to another. Once it is, the molecule absorbs energy into this form as freely as any other.

Solids

What should molar heat capacities be for solids? Let’s just think about a simple solid, made of a regular lattice of just one kind of atom. Each atom in the solid can vibrate along x, y, or z. Each of these vibrations involves both kinetic and potential energy, so there are $2 \times 3 = 6$ degrees of freedom. This would lead us to expect $C_{\text{solid}} = \frac{6}{2}R = 3R$, or around 25 J/mole*K, which is in fact what you usually find.

What would the molar heat capacity be for some more complex material, like ice, a solid made of molecules of water? For this, you might have the six modes seen for simple solids, in which the whole molecule oscillates back and forth in each direction, then also have modes for bending of the molecule, or oscillation of the molecule around different axes, or vibration along the bonds in the molecule. Each of these might, if it is accessible (doesn’t have quantum states that are too far apart) add to the specific heat. So we’d certainly expect C_{ice} to be bigger than $C_{\text{simple solid}}$. And again, that’s just what we find. The molar heat capacity of ice is actually quite high; about 75 J/mole*K.

Molar heat capacity and specific heat

Molar heat capacity is the fundamental thing, as it tracks how much energy goes into each atom or molecule. But in ordinary life we don’t count the number of atoms in a sample, we weigh it. So it is useful to know how much heat is required to raise the temperature of a certain *mass* of material. This is the specific heat:

$$\Delta Q = mc_{\text{specific}} \Delta T$$

From this we can see that $c_{\text{specific}} = C_{\text{molar}} / M_{\text{molar}}$. The fact that different materials have different molar masses means that, although they may have the same molar heat capacities, they will have quite different specific heats. Substances like the noble gases vary in molar mass by a factor of 30, so their specific heats will also vary by a factor of 30. Typical values for specific heats range from 500 to 15,000 J/kg*K. Usually they’re tabulated as J/gram*K, so the numbers you might see in a table listing these would be 1000 times smaller.

In general substances with massive atoms will have low specific heats, though only because a sample of fixed mass is actually made of fewer atoms.

Heat capacity and energy storage

Once we know something about heat capacity, we can learn an important basic fact about thermal energy; we can quantify it. Let's think about an example. Imagine you have a boiling hot cup of coffee. Eventually, it will cool off to room temperature, releasing some of its thermal energy into its surroundings. How much energy will this be?

Coffee is mostly water. We know that the specific heat of water is around $4200 \text{ J/kg} \times \text{K}$. An 8 oz cup of water has a volume of about 240 cm^3 or $2.4 \times 10^{-4} \text{ m}^3$ and a mass of about 0.24 kg . When it cools from boiling (373 K) to room temperature (298 K), its temperature drops by about 75 K. The total energy released is $(4200 \text{ J/kg} \times \text{K})(0.24 \text{ kg})(75 \text{ K}) = 75,600 \text{ Joules}$. Is that a lot of energy?

If you could take this much energy and convert it into kinetic energy in your body, you would end up traveling at:

$$\frac{1}{2}mv^2 = \frac{1}{2}(80 \text{ kg})v^2 = 75,600 \text{ Joules}$$

or

$$v = 43.5 \text{ m/s} = 97.4 \text{ mph}$$

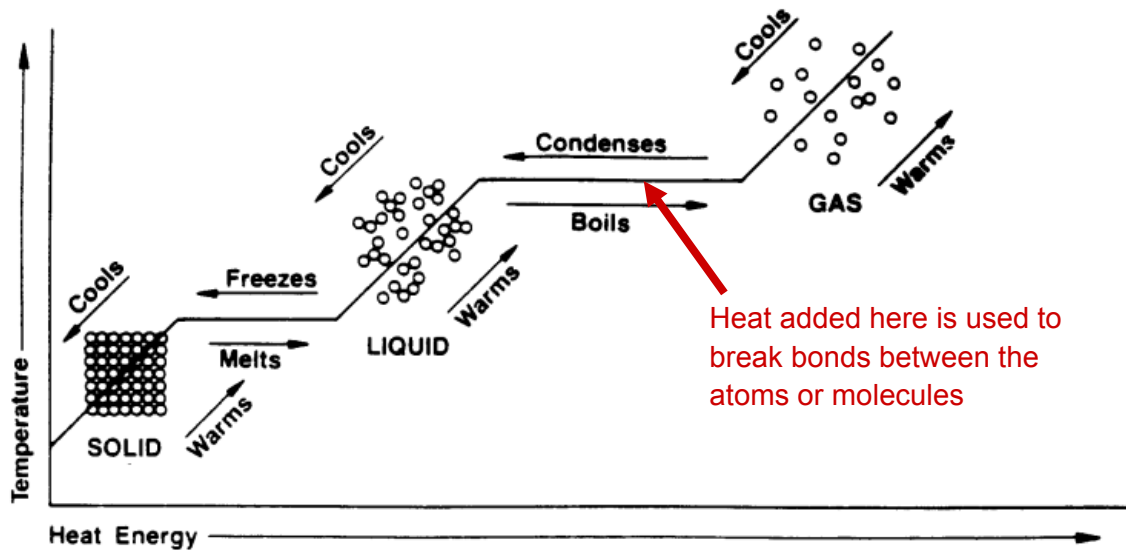
So if you could somehow extract the excess energy in that cup of hot coffee and use it to *really* get yourself going in the morning, you could in principle shoot out of the kitchen at 97 miles per hour. And that's without even considering the caffeine! Not surprisingly, the proposed transformation (take all the excess heat and organize it into your motion) is impossible. We will see why more clearly when we talk about heat engines below.

Although this is kind of a joke example, it does get at an essential point. There is a LOT of energy around in thermal form. This should not be a huge surprise, as we already know of the tendency for energy in any other form to quickly spread out into thermal energy. But it is remarkable just how much energy there is in something like a cup of coffee.

Changes of state and heat of transformation

There is one more thing to discuss when talking about how the addition of heat to materials causes changes of temperature; changes of state. When a material goes through a change of state from solid to liquid or liquid to gas, additional heat is required to break the bonds which are holding atoms together. This is illustrated in the figure below.

During the period while the change of state is occurring, heat which is added does not raise the temperature, but rather goes into breaking the bonds that hold the atoms together. The heat required to do this is quantified as the "latent heat of transformation", and written $\Delta Q = mL$. So to transform a mass m from solid to liquid, without raising the temperature, you get it to the melting point then add an amount of heat equal to the mass times the latent heat to break all the bonds.



These latent heats have a wide range of values, from 10^4 J/kg*K to 2×10^6 J/kg*K. So they are large, larger than specific heats. It takes more energy to break the bonds in a material than to raise the temperature of that same material by one degree.

Tables giving values for some specific heats, molar heat capacities, and latent heats of transformation are given on the next page.

There are a few features of these specific and latent heats which are obviously important for life. As always, water is unusual in a useful way. Its high specific heat helps it to play a role in temperature stability. If you're a 60 kg person, made of mostly water, it takes around a quarter of a million joules to change your temperature by 1 K.

Similarly the high latent heat for freezing water makes it difficult to freeze a pond all the way down to the bottom, making it easier for life to survive there.

Substance	Phase	c_p $\text{J g}^{-1} \text{K}^{-1}$	C_p $\text{J mol}^{-1} \text{K}^{-1}$
Air (Sea level, dry, 0 °C)	gas	1.0035	29.07
Air (typical room conditionsA)	gas	1.012	29.19
Aluminum	solid	0.897	24.2
Ammonia	liquid	4.700	80.08
Argon	gas	0.5203	20.7862
Copper	solid	0.385	24.47
Diamond	solid	0.5091	6.115
Gold	solid	0.1291	25.42
Helium	gas	5.1932	20.7862
Hydrogen	gas	14.30	28.82
Iron	solid	0.450	25.1
Mercury	liquid	0.1395	27.98
Nitrogen	gas	1.040	29.12
Oxygen	gas	0.918	29.38
Silica (fused)	solid	0.703	42.2
Uranium	solid	0.116	27.7
Water	gas (100 °C)	2.080	37.47
	liquid (25 °C)	4.1813	75.327
	solid (0 °C)	2.114	38.09
All measurements are at 25 °C unless otherwise noted. Notable minimums and maximums are shown in bold.			

Substance	Latent Heat Melting J/g	Melting Temp °C	Latent Heat Vaporization J/g	Boiling Temp °C
Alcohol, <i>ethyl</i>	108	-114	855	78.3
Ammonia	339	-75	1369	-33.34
Carbon Dioxide	184	-57	574	-78
Helium			21	-268.93
Hydrogen	58	-259	455	-253
Nitrogen	25.7	-210	200	-196
Oxygen	13.9	-219	213	-183
Water	335	0	2272	100

15.4 Moving heat around

We know that if we put hot and cold things near one another, thermal energy will flow around until an equilibrium temperature is reached. There are three principle methods by which this heat will move

- Conduction
- Convection
- Radiation

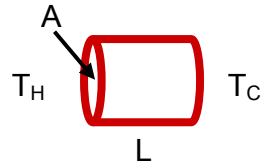
We will discuss each in turn. All operate within the basic picture of statistical physics, which suggests that heat will tend to spread outward from places where it is localized.

Conduction

Conduction is the form of heat transfer which occurs through direct contact. For example, if you put one side of a slab in contact with something hot, and the other side in contact with something cold, heat will flow directly through the slab, as the atoms on the hot side rattle off their neighbors, which bang into their neighbors, and so on until the atoms on the cold side are heated up.

$$\frac{dQ}{dt} = -kA \frac{dT}{dx}$$

$$\frac{dQ}{dt} = -kA \frac{T_H - T_C}{L}$$



The flow of heat by conduction depends first on some obvious geometric factors. If the slab is thick, heat will flow more slowly. If the slab has a large area through which heat can flow, heat will flow more rapidly. Finally, the rate at which heat flows depends on the nature of the material itself, through a parameter called the *thermal conductivity* k of the material. These are combined together in the following way:

The first equation shows that, at any point, the rate of heat flow will depend on the thermal conductivity, the area through which it can flow, and the temperature gradient. The second equation shows what this would be for a cylindrical slab like of area A and length L , which has a hot end (T_H) and a cold end (T_C). The thermal conductivity k has units J/msK , or since $1 J/s = 1 \text{ Watt}$, you can write this as W/mK . Here is a table of some thermal conductivities:

Material	Thermal conductivity ($W \cdot m^{-1} \cdot K^{-1}$)	Temperature (K)
Copper (Cu), pure	385 - 401	273-373
Gold (Au), pure	314 - 318	273-373
Aluminum (Al), pure	205 - 237 (220)	293
Carbon Steel (Fe+(1.5-0.5)%C)	36 - 54	
Ice	1.6 - 2.2	273
Water	0.6	293
Wood	0.16 - 0.4	298-293
Snow (dry)	0.11	
Air (1 atm)	0.024 - 0.0262	273-300

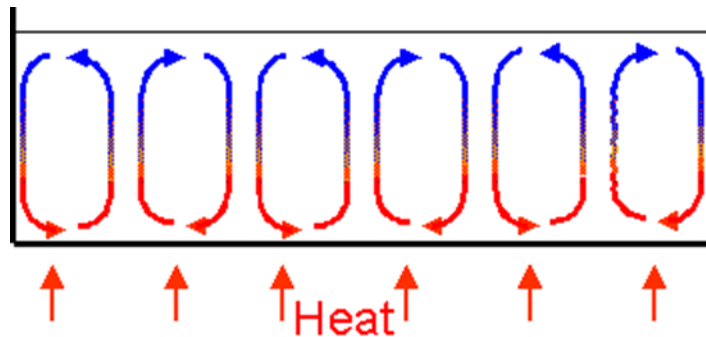
There are some things to notice about this. First, the metals at the top have very high thermal conductivity. Heat flows through these things very easily. This fact is related to their *electrical conductivity*. This high conductivity is why we use primarily metal pots and pans, rather than ceramic for

instance. The other solids, like ice and wood, conduct heat hundreds of times more slowly. They make nice insulators. Better still is air, which conducts heat 15,000 times more slowly than copper. Air can make a great insulator.

Convection

A second way to move heat around is called convection. In convection you move heat simply by mixing hot and cold material. Mostly this happens in liquids and gases, where flow and the consequent mixing of material is easy. You cause heat flow like this when you stir cream into your coffee. This is sometimes called forced convection. A more important process is “free convection” which uses gravity to mix the hot and cold material, so that the process happens without intervention.

Since most materials expand when they are heated, hot material is usually less dense than cold. If a fluid is pulled on by gravity, this less dense material will rise due to the buoyant force, which we will explore in more detail later in the class. Any time you have a fluid which is hotter on the bottom than on the top, convection will work to mix it. One case you’re familiar with is heating a pot from the bottom.



Since the only life we know is on Earth, and it all lives under the influence of gravity, free convection like this is a very important process in nature. It drives atmospheric and oceanic currents, and is responsible for plate tectonics.

Radiation

The third process of heat transfer is due to electromagnetic radiation. The familiar form of electromagnetic radiation is visible light, but it appears in many guises. They are all really the same, and differ only in wavelength. Other forms of electromagnetic radiation include radio waves, microwaves, infrared radiation, ultraviolet radiation, x-rays, and gamma-rays.

We have learned that the atoms which make up all materials are always in motion, shaking back and forth and bouncing off one another. It turns out that any time a charged particle (like an electron or a nucleus) is accelerated, it will emit electromagnetic radiation. The thermal motion of atoms always involves these accelerations, and they’re all made of charged particles, so *all* materials are *always* emitting electromagnetic radiation, and this radiation carries away energy.

The amount and nature of the radiation each object emits depends primarily just on temperature. The basic rule is encoded in Stefan’s Law:

$$\frac{dQ}{dt dA} = R = e\sigma T^4$$

The heat loss per unit area per unit time R depends on the temperature to the fourth power, a universal constant σ , and a property of the surface emitting the radiation called the emissivity e . The constant σ , called Stefan's constant, has a value of:

$$\sigma = 5.7 \times 10^{-8} \frac{\text{W}}{\text{m}^2 \text{K}^4}$$

The emissivity is a measure of how good a radiator this surface is. It depends on both the material the object is made of and the condition of its surface (is it rough or smooth, clean or contaminated). The emissivity takes on values from zero to one only. A perfect emitter would have $e = 1$ and a completely failed emitter would have $e = 0$.

Notice that the heat loss rate changes very rapidly with the temperature, as the fourth power. If you double the temperature, you increase the heat loss rate by sixteen times. What's the typical scale for this? Assume $e = 1$, and room temperature, around 300 K. For this you get a heat loss rate $R = 460 \text{ W/m}^2$. Every object at room temperature emits approximately 460 Joules per second from every square meter of surface it has.

This is quite a loss rate. If everything is radiating away energy like this, why doesn't everything rapidly cool off? The reason objects don't cool wildly is because radiation goes both ways. A book on the table (for example) is surrounded by other objects at roughly the same temperature. Those other objects are emitting radiation which is then absorbed by the book. So while radiation is going out of the book, it also comes in from its surroundings. Notice that this balancing process (out and in) will tend to make all the objects approach the same temperature.

There is another important point here. The emission of radiation and the absorption of radiation are inverse processes. That is to say, they're really the same, just reversed in time. Because of this, any material which is a good emitter ($e \approx 1$) is also a good absorber, and any material which is a bad emitter ($e \approx 0$) is a bad absorber. This fact aids in allowing radiation to balance temperature. Anything which emits a lot will also absorb a lot, while anything which emits very little will absorb very little.

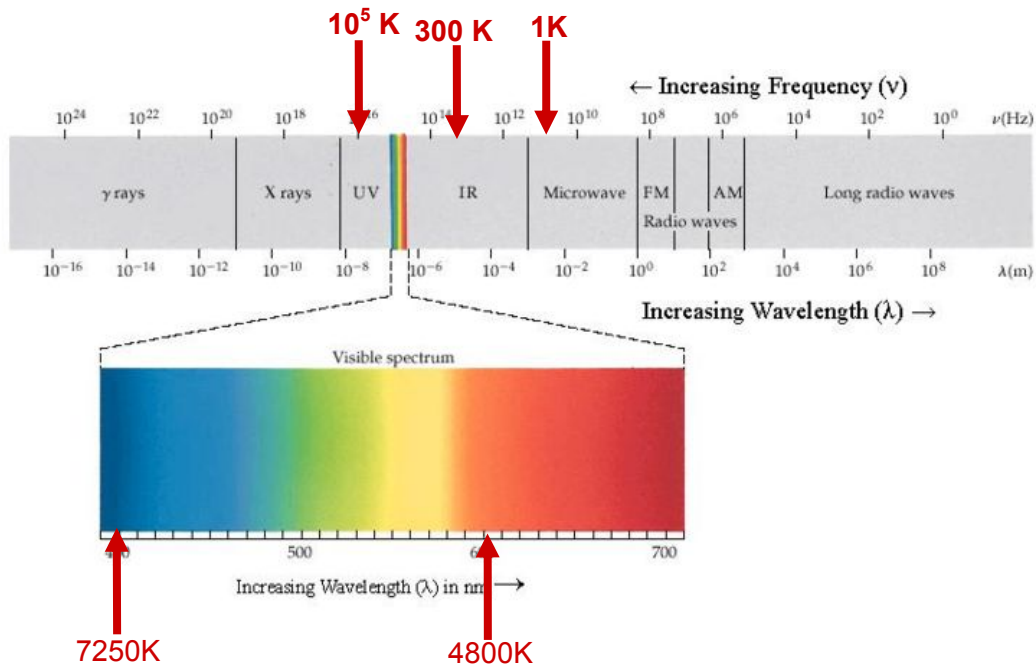
Think a bit about this. Which is a better emitter of visible light, a piece of black paper or a mirror?

The nature of thermal radiation

If everything is emitting radiation all the time, why can't you see in the dark? Well, you can see some things in the dark, things which are hot enough to emit visible light! The thermal radiation emitted by objects comes out in a broad distribution of wavelengths, but the peak of this distribution changes with temperature according to Wein's law:

$$\lambda_{\text{peak}} = \frac{2.9 \times 10^{-3} \text{ mK}}{T}$$

For a temperature of 300 K, this is about 10 μm . Radiation with this wavelength is in the infrared part of the spectrum, a part which your eyes cannot detect. As a result, you don't see room temperature things through light which they emit. You only see them through visible light which they reflect. The figure below shows the "electromagnetic spectrum" and points out what part of the spectrum is emitted by objects of different temperatures. To emit significant amounts of visible light, an object must reach around 1000 K or so.



15.5 Size, thermoregulation, and life

Life uses random thermal motion to accomplish many things, including things like diffusion and osmosis. One strategy for ensuring the stability of these thermal effects is to maintain a constant temperature. At constant temperature chemical and physical processes run along at a fixed rate. An organism which can stabilize its temperature can count on high performance at all times.

This constant temperature approach, taken by "homeotherms", is effective, but very costly in terms of energy usage and control complexity. Maintaining a constant temperature also requires constant vigilance. Many of our functions fail catastrophically if body temperature moves outside relatively narrow limits. The other major approach is taken by "ectotherms", which generally take on the temperature of their surroundings. They don't have to pay the substantial metabolic cost of maintaining their temperature, but they must survive with performance which varies strongly with temperature.

Given the variety of mechanisms for heat loss, maintaining a constant temperature comes at significant cost. People, for example, must maintain their temperature in quite a narrow range around 37° C. If your temperature falls below 32° C you would suffer severe hypothermia and after a short time would die. If

your temperature rises above about 41° C you may suffer brain damage, and above 45° C death is nearly certain.

Why do these dire consequences occur? At low temperatures, your normal cellular processes slow down, eventually becoming too slow to maintain the finely coordinated sequence of reactions that your body requires. At high temperatures everything cooks along, but the temperature becomes high enough to begin to destroy some of the very proteins you need for your cells to function. This narrow survival range is one of the clearest signs of how important these purely random thermal processes are for life.

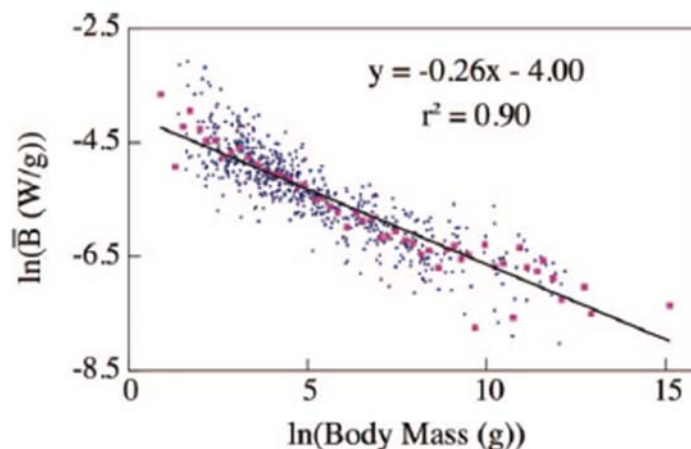
Size is a major factor in the challenge of thermoregulation. Recall that surface area varies like size², while volume varies like size³. The generation of heat in an organism depends on the number of cells, and so is proportional to volume of an organism. Heat loss, by whatever means, depends on surface area. So the ratio of heat lost to heat generated changes as:

$$\frac{\text{heat lost}}{\text{heat generated}} \propto \frac{\text{size}^2}{\text{size}^3} = \frac{1}{\text{size}}$$

When organisms are large, losing too much heat is not a problem. In fact, for very large organisms, the challenge lies in getting rid of the really substantial heat they produce. Large organisms must be very careful not to overheat. This is why the really large mammals are often hairless and include specialized heat loss mechanisms like the very large ears of elephants (which have very large surface area and little volume).

On the other end of the size scale, small creatures lose the heat they produce very quickly. At this extreme lie creatures like shrews and hummingbirds. To replace the heat they lose so rapidly these tiny creatures must consume relatively enormous quantities of food, many times their own weight each day. Even doing this, these creatures live on the edge, and unlike many homeotherms must be able to survive dramatic reductions in body temperature. When faced with suddenly cold conditions they can enter a dormant state, allowing their body temperature to drop. Provided they can eventually use environmental heat to recover (sunlight for example) they survive this without harm. Many small homeotherms follow this strategy.

It has long been known that the basal metabolic rate of mammals generally decreases with size. Small ones like shrews use more energy per gram per second than large ones like elephants. This is shown in the figure below, drawn from a recent paper on the subject¹. This figure shows that the mass specific metabolic rate generally decreases with body mass. While there is substantial variation at any particular



body mass, the overall scaling is clear. While it is unlikely that this scaling law is solely determined by thermoregulation concerns, it seems clear that they play a role in its nature.

Methods of controlling heat loss and gain

Most homeotherms have temperatures which are elevated relative to their surroundings. Most of the time, they will want to prevent heat loss in an effort to reduce the metabolic cost of maintaining a high temperature. But occasionally they will find themselves in circumstances where they begin to overheat and have to dump excess heat. Because so much of life operates on thermal energy, this is an important problem, and solutions to it are extremely various.

The thermal challenges facing life are very different for organisms living in air and water. Terrestrial life faces enormous variation in environmental temperature and radiative environment. That these variations take place on many time scales (the change of seasons, the daily cycle, the passage of a cloud over the sun) only makes the problem more challenging. Living in water alters, and in some ways moderates, the thermal challenges organisms face.

All organisms, even homeotherms, allow some variation in temperature within their bodies. In particular surface temperature is often much lower than core temperature. So while external loss of heat from the surface is the fundamental problem, we will have to consider how organisms minimize the internal loss of heat from the core to the surface as well. In humans, for example, core temperature is typically 37° C while surface temperature is more like 33° C. So there are both internal and external temperature gradients to consider.

Every organism lives immersed in either air or water, so conductive heat loss through these materials represents the minimum thermal connection an organism can have with its environment. The thermal conductivities of air and water are quite different:

$$k_{\text{air}} = 0.024 \frac{\text{W}}{\text{mK}}$$
$$k_{\text{water}} = 0.59 \frac{\text{W}}{\text{mK}}$$

The thermal conductivity of air is about 25 times lower than that of water. This fact alone ensures that all aquatic organisms will be in much closer thermal contact with their environment than those which live on land. Indeed most organisms which live in water are ectotherms whose body temperatures largely match the temperature of their surroundings. Fascinating exceptions exist, but the high thermal conductivity of water requires them to take extreme measures to prevent heat loss. Since the thermal environment on land and in the sea differs so strongly, we will consider them in turn.

Thermal regulation on land

Most of the time, homeotherms need to maintain a body temperature elevated above their surroundings. Doing this requires limiting heat loss. Any organism with a surface temperature different from its surroundings may lose heat through conduction, convection, and radiation. So the first step is often to allow the surface temperature to match, or nearly match, the surroundings. This is the function of

insulation of many kinds. Mammals use fur, birds use feathers; the mechanism is much the same. By trapping air and preventing convection in it, fur and feathers can achieve thermal conductivities which essentially match the low conductivity of air. Both fur and feathers are adaptable as well. You have

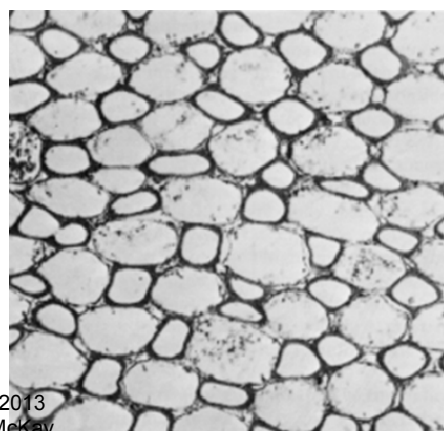


probably seen squirrels and birds fluffed up in the winter. Each is thickening the layer of low conductivity insulation surrounding it. In mammals fluffing up your coat is called piloerection, and your own attempts at this (goosebumps) don't help much.

Fur and feathers are all about preventing heat loss from the skin surface. It works very well, and allows most mammals and birds to have skin temperatures which differ little from their core temperatures; external insulation is enough.

There are cases, even for mammals and birds, where external insulation is not enough to protect a core elevated in temperature. One example is the bare legs of water birds. Watching ducks swim in icy ponds makes one wonder how they do it. They don't seem to mind at all. Their secret, one shared by an extraordinary variety of organisms, is the counter-current heat exchanger.

This remarkable structure is an area in which arteries carrying warm oxygenated blood from the core of the organism are divided into many small vessels which are then intertwined with similarly fragmented veins bringing cold, oxygen poor blood back from the ducks feet. When this network of arteries and veins intertwines, it allows the transfer of heat coming out of the body into cold blood coming back returning. In this way, the heat which would flow out and be lost to the environment is instead restored to the core. An example, in this case from a Skipjack Tuna is shown below. In it, the arteries are the smaller, thickwalled vessels, while the veins are the larger ones. The tight intertwining of the two is apparent.



The use of this mechanism is very widespread in nature, often at the bases of appendages like the tails of muskrats and beavers, and the legs of leatherback turtles. Such a mechanism helps to prevent overheating in the brains of ungulates like sheep and in the testes of many mammals, including humans. Even insects like honeybees do this, using it to insulate their thorax, where wing muscles generate substantial heat, from their abdomens. Not surprisingly human technology has adopted this nice mechanism, and a wide variety of flow oriented heat exchanges are used.

Thermal regulation in water

The much higher thermal conductivity of water makes the challenge of homeothermy greater in aquatic environments. Most aquatic organisms are ectotherms, simply adopting the temperature of the environment. But the advantages of elevated temperature, particularly for predators, are large enough to drive evolutionary innovation.

The main groups of homeotherms living in water are mammals and birds which have returned to the sea. In them, evolution has found solutions to the thermal challenges of life in water. In mammals, one key element was the development of blubber. Blubber has many functionsⁱⁱ, streamlining the body, providing an energy storage site, and contributing to buoyancy. But perhaps its most important purpose is internal insulation. Blubber has a thermal conductivity substantial less than water, about 0.15 W/mK , about a factor of four lower than water. Thick layers of this enable marine mammals from seals to whales to comfortably feed in the rich waters of the Arctic and Antarctic oceans. Seals can have skin temperatures about 35°C lower than their core temperatures with no discomfort. Blubber does not completely cover most marine mammals; fins for example remain exposed. Perhaps not surprisingly, heat loss to these is limited by counter-current heat exchangers.

Perhaps the greatest challenge for heat loss is respiration. Oxygen is acquired by diffusion, which requires intimate contact between air (or water) and a very large surface area containing blood. This large area of close contact insures transfer of heat as well as oxygen. The air you breathe in on a cold winter day reemerges at a temperature close to your core temperature. So you must pay the thermal cost of heating that air with every breath. In water, extracting oxygen is even more costly. The water in close contact with the blood in gills has higher conductivity, higher heat capacity, and higher density than air. An organism which breathes water will inevitable lose much more heat in respiration than one which breathes air. We should not be surprised that almost all aquatic homeotherms breathe air. It is no doubt annoying to have to return to the surface to breathe, but it does avoid the otherwise inevitable heat loss associated with respiration.

A Quick Summary of Some Important Relations

Thermal expansion:

Most solids and liquids expand when heated. This can be modeled for linear and volume expansion as:

$$\Delta L = \alpha L \Delta T \quad \text{and} \quad \Delta V = \beta V \Delta T \quad \text{with} \quad \beta \approx 3\alpha$$

Basics of heat capacity:

To change the temperature of some matter, you might add some heat:

$$\Delta Q = mc_{\text{specific}} \Delta T = nc_{\text{molar}} \Delta T$$
$$c_{\text{molar}} = M_{\text{molar}} c_{\text{specific}}$$

Predicting heat capacities:

To raise the temperature of some matter, you must increase the average KE of its atoms. Energy put into such matter will spread equally into every form which it might take (simply because this is most likely). So the amount of energy you must put in to change the temperature depends on how many bits of energy you must put in. Each atom must receive energy for each ‘degree of freedom’; each way it can store energy. Each DOF will receive $\frac{1}{2}k_B\Delta T$ for each atom. This leads to molar specific heats which are:

$$c_{\text{molar}}^{\text{monatomic gas}} = \frac{3}{2}R$$
$$c_{\text{molar}}^{\text{diatomic gas}} = \frac{3}{2}R \text{ or } \frac{5}{2}R \text{ or } \frac{7}{2}R \text{ (typically } \frac{5}{2}R \text{ at room temp)}$$
$$c_{\text{molar}}^{\text{monatomic solid}} = \frac{6}{2}R$$

The changes in diatomic specific heat with temperature reflect the impact of quantum mechanics on these molecules. Until the average thermal energies are large enough to ‘excite’ rotation and vibration, they cannot absorb energy.

Changing state and latent heat:

The heat needed to change state is described by ‘latent heat’ for melting or vaporization.

$$\Delta Q = mL$$

Heat flow by conduction:

This is governed by the thermal conductivity of the material k and the temperature gradient.

$$\frac{dQ}{dt} = kA \frac{dT}{dx}$$

Heat flow by convection:

Convection is mathematically complex: no especially useful simple formulae are available.

Heat flow by radiation:

All objects radiate away energy in the form of electromagnetic radiation constantly. The manner in which they do that is governed by just a few universal relations:

$$R = \frac{dQ}{dt dA} = e\sigma T^4$$
$$\lambda_{\max} = \frac{2.838 \times 10^{-3} \text{ m/K}}{T}$$

Scaling laws and thermoregulation:

Heat is generated throughout the volume of an organism, and lost through its surface. This makes too much heat loss a problem for small organisms and overheating a problem for large organisms.

Thermoregulation in air and water:

Maintaining an elevated temperature is much more difficult in water than air, both because of water's high thermal conductivity and heat capacity. This challenge is especially difficult for organisms that extract oxygen from water by diffusion.

ⁱ Savage, V., *et al.*, 2007, "Scaling of number, size, and metabolic rate of cells with body size in mammals", Proceedings of the National Academy of Sciences, **104**, 4718.

ⁱⁱ Dunkin, R., *et al.*, 2005 "The ontogenetic changes in the thermal properties of blubber from Atlantic bottlenose dolphin *Tursiops truncatus*", Journal of Experimental Biology, **208**, 1469.

16. Structures and processes in a world of randomness

- 1) Forming structures using random motion
 - i. Examples from the non-living world: snowflakes
 - ii. Examples from the non-living world: planets
 - iii. **Examples from the living world: DNA strands**
- 2) Processes in the non-living world
 - i. Convection and its children: hurricanes, tornados, and the water cycle
- 3) What is life?
 - i. A multitude of definitions
 - ii. Lack of clear consensus
 - iii. Life and motive force
- 4) Thermodynamic cycles and engines: motive force from technology
 - i. Expansion and contraction of a gas: work done
 - ii. Isothermal expansion and contraction
 - iii. Constructing the Stirling cycle as a tidy example
- 5) Efficiency in thermodynamic cycles
- 6) How life works: isothermal circumstances and free energy

Physics for the Life Sciences: Chapter 16

16.1 Forming structures using random motion

We've been learning about how life makes use of random thermal motion to get things done. The practical inevitability of the increase of entropy drives mechanisms like diffusion. We have seen that to make these processes happen reliably, many organisms maintain a fixed, generally elevated temperature. In simple closed system, this random motion tends to an equilibrium which is simple and boring; with all the matter and energy spread out more or less uniformly. Indeed that's what will happen to you if we seal you off from the world. Pretty quickly you will end up at in an unchanging (dead) equilibrium state. Once there, nothing would every happen again.

Life does something much more interesting than this. Somehow it manages to construct from its environment complex, active structures which grow, interact with the world, and survive for long periods of time. This seems to fly in the face of the inevitable increase in entropy. The ability of life to do this with such wild inventiveness is the source of much of its magic. How can life harness randomness and still form complex, interesting structures? How can such structures form unbidden in a universe governed by the steady increase of entropy?

Examples from the non-living world: snowflakes

The apparent conflict between the second law and the formation of structure is resolved by recalling that most of the systems we have studied were carefully arranged as *closed* systems. Typically neither matter nor energy is allowed to enter or leave. Such a system, like our box filled with gas atoms, will indeed rapidly approach a uniform, structureless, thermal equilibrium. Once it does so nothing significant will ever happen. Atoms will still move around, things change on the microscopic level, but the macroscopic state of the system will never change.

But few systems on Earth are ever so carefully isolated. Energy flows into and out of systems constantly. This changes the accounting completely, and allows surprisingly complex structures to form. To see how this works, leave life aside for a bit and consider how something as impressive as a snowflake forms.

It begins with some water vapor, spread more or less at random, containing some energy. Surround this vapor with a cold environment and the energy within it will, as usual, spread out as much as possible. This spreading could happen by just cooling the vapor, taking away the energy of each atom leaving them separate.

But it's possible to do more. Water molecules can give up even more energy if they lock together in a solid form. To release the maximum amount of energy, each molecule must lock into a very regular crystal structure with its neighbors. When it does this, even more of the energy originally localized in the water is allowed to spread out; an outcome which statistical mechanics suggests is inevitably likely.

The incredible structures you see in snowflakes are all examples of how such crystals can form when they grow. The great variety of structures you see come about because the conditions for crystal growth are each time just a bit different. In a laboratory it is possible to make the conditions very stable, and hence to repeatedly grow the same kinds of snowflakes. These amazing structures form completely by chance, through the operation of the same thermodynamic laws which guarantee that gas atoms released in a box will spread uniformly through it. They do so because, although the final structure of the atoms in these snowflakes is extremely ordered, the final distribution of the energy which they once possessed is vastly more disordered. Taken together, entropy has increased.



You can see rich and beautiful examples of this snowflake growth at the research page of Caltech Physicist Ken Libbrecht.¹

Examples from the non-living world: planets

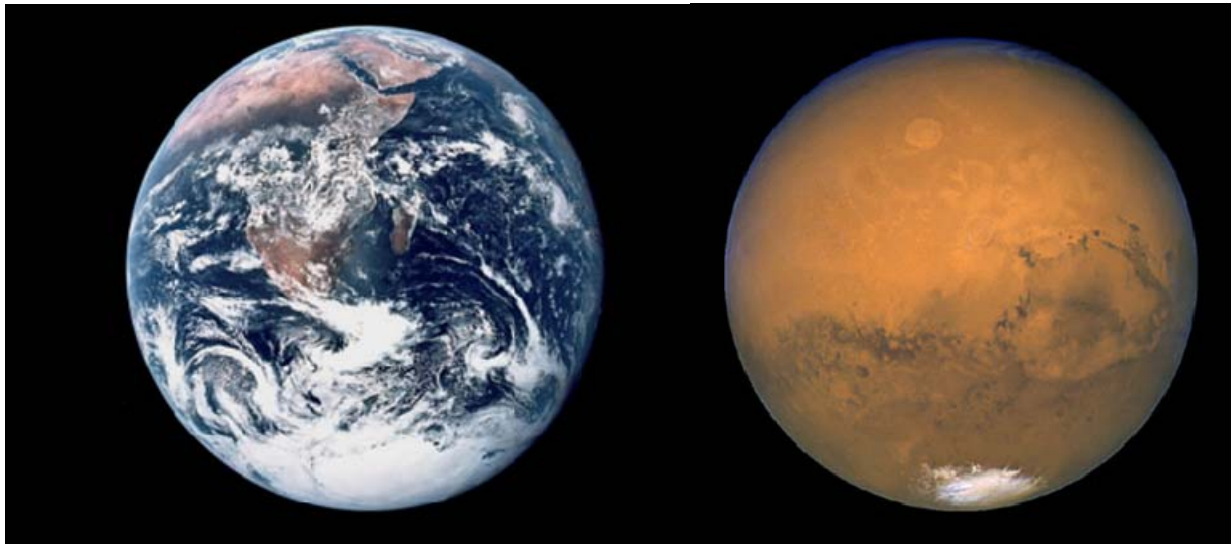
Another fine (though perhaps less exquisite) example of unlikely structure formation are the planets. Planets are remarkably spherical. The Earth, for example, has a radius of 6.38×10^6 m. The largest features

¹ <http://www.its.caltech.edu/~atomic/snowcrystals/>

on its surface like Mount Everest (8.85×10^3 m high) and the Marianas Trench (10.9×10^3 m deep) are less than 0.2% of its radius. It's really smooth. Now it isn't perfectly spherical. Since it spins on its axis, it actually bulges outward at the equator. The distance from the Earth's center to its surface is a little over 0.3% bigger at the equator than it is at the poles. Let's put it this way. If you had a ball this spherical, you'd be pleased. How do they get so round?

Planets form by gravitational accretion. Imagine some region in space with a lot of dust, gas, and rocks around. Since this stuff is initially far apart, each pair of things has a lot of gravitational potential energy. Any region with density a little high compared to the average has adequate mass to pull in neighboring material, converting its excess gravitational energy into kinetic energy as it falls in. Once things collide, they typically convert the bulk kinetic energy they have gained into random thermal motion. This thermal motion is then usually radiated away, leaving the material forever and spreading out through space, increasing the overall entropy of the universe.

The more closely together you pack in this infalling stuff, the more energy you can release, and the larger the increase in entropy will be. The best way to pack it in is to make the final object a sphere. This spherical shape releases much more energy, increasing the final entropy a lot. For this reason, planets of a decent size will always become spherical.



Structure formation in the non-living world is all around you, including the formation of galaxies, stars, and planets, all the way down to diamonds, raindrops, and clouds. These structures form because in doing so entropy increases.

The same essential mechanisms are used by life for the creation of complex structures. When your cells build a protein, they don't grab each atom, carry it across the cell and carefully put it in place. Instead, they just have the right ingredients in the right conditions to allow random thermal motion to put the protein together. Making those conditions just right for all the interlocking constructs of biochemistry to appear in a coordinated way is what makes life seem so different from non-living matter. But in fact the same essential processes are occurring.

16.2 Processes and cycles in the non-living world

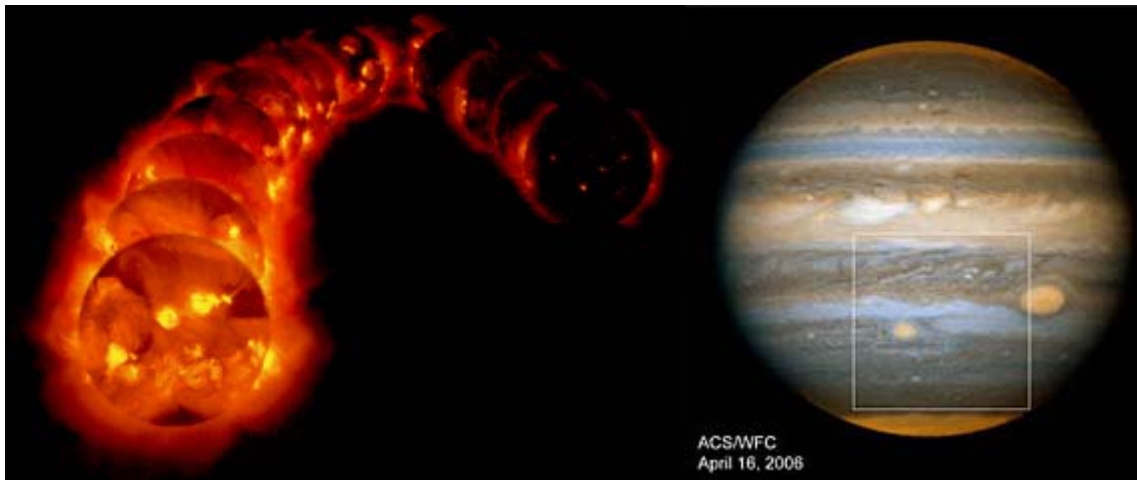
The formation of structures along the march toward increased entropy can be understood. But life is more than the formation of structure. Living things are active, they move around, grow, do things repeatedly rather than once. Living things have many cycles. How can cyclic processes like this occur if the inexorable march toward high entropy is behind everything that's happening?

Again, the problem is in the accounting. Closed systems, left alone, can't undergo cyclic behavior. Somewhere along the way this would have to violate the second law of thermodynamics. But the Earth is hardly a closed system. Energy flows into the Earth continuously from the Sun. It flows out continuously as well, spreading out into space. Take away the Sun and the Earth's surface would rapidly cool to the nearly the background temperature of space, about 3 K. Stop the Earth from radiating away heat and it would rapidly overheat. Of course that's perhaps what we're doing right now...

So the Earth is a system through which energy flows. When energy flows through a system, cyclic processes emerge very naturally. Think a bit about what the Earth would be like without life. Seen from space it would seem very much the same. Weather cycles would continue, cycling water through the atmosphere, lifting it across the continents where it could flow back toward the oceans. The annual weather cycles, driven by the changing orientation of the Earth relative to the sun would be unchanged. Thunderstorms, hurricanes, and tornados would continue to naturally concentrate enormous amounts of energy as a consequence of accidental conditions. Ocean currents would flow, annually carrying icebergs to lower latitude. Volcanoes would erupt, reforming the land, while within the Earth convection in the mantle would continue to drive continental drift, rearranging the continents every few hundred million years. The Earth, without life, remain a very lively place.

There are other places in the solar system where similarly interesting cyclic, grossly repeated behavior happens on its own. The Sun, site of such a huge energy flow, is perhaps most dramatic, with wild surface storms rising and falling in number through an ~ 11 year cycle. Similar weather affects the other planets too, perhaps most impressively on Jupiter, where the red spot (a 400 year old storm) was joined a few years ago by a new storm.

Most of these phenomena are driven by convection. They "just happen" because they allow energy to spread more thoroughly, speeding its escape from these open systems.



Two pictures of lively weather elsewhere in the solar system: On the left is a montage of images of the Sun taken as it progress from the peak of the solar cycle (on the left) to the minimum. On the right is a picture of Jupiter from April 2206, showing the new storm which has recently joined the famous “red spot” storm first observed by Giovanni Cassini in 1655.

16.3 What about life?

Is life just a kind of natural thermodynamic process which happens because it speeds the increase of entropy? To think about this more, it's useful to ask what life is, where we draw the line between the living and the non-living. There is no complete consensus about what marks this line. The Oxford English Dictionary waffles, defining life as:

“The property which constitutes the essential difference between a living animal or plant, or a living portion of organic tissue, and dead or non-living matter; the assemblage of the functional activities by which the presence of this property is manifested.²”

I don't know about you, but I don't find this very satisfying. This definition says life is what makes the difference between living and non-living matter. Ernst Mayr, one of the 20th century's leading biologists, expressed his exasperation with the problem this way: “Attempts have been made again and again to define 'life'. These endeavors are rather futile since it is now quite clear that there is no special substance, object, or force that can be identified with life.³”

Still it seems useful to try. Probably the narrowest, most widely accepted definition of life is something which reproduces itself with the possibility of modification. This lets in all the widely accepted living things. Some prefer a longer list of criteria, here's a version drawn directly from the Wikipedia⁴ as of 2007, which captures most of the properties usually raised:

² Oxford English Dictionary online

³ Mayr, E. 1982. *The Growth of Biological Thought. Diversity, Evolution, and Inheritance*. Harvard University, Cambridge: The Belknap Press, .

⁴ <http://en.wikipedia.org/wiki/Life>

1. **Homeostasis:** Regulation of the internal environment to maintain a constant state; for example, sweating to reduce temperature.
2. **Organization:** Being composed of one or more cells, which are the basic units of life.
3. **Metabolism:** Consumption of energy by converting nonliving material into cellular components (anabolism) and decomposing organic matter (catabolism). Living things require energy to maintain internal organization (homeostasis) and to produce the other phenomena associated with life.
4. **Growth:** Maintenance of a higher rate of synthesis than catalysis. A growing organism increases in size in all of its parts, rather than simply accumulating matter. The particular species begins to multiply and expand as the evolution continues to flourish.
5. **Adaptation:** The ability to change over a period of time in response to the environment. This ability is fundamental to the process of evolution and is determined by the organism's heredity as well as the composition of metabolized substances, and external factors present.
6. **Response to stimuli:** A response can take many forms, from the contraction of a unicellular organism when touched to complex reactions involving all the senses of higher animals. A response is often expressed by motion, for example, the leaves of a plant turning toward the sun or an animal chasing its prey.
7. **Reproduction:** The ability to produce new organisms. Reproduction can be the division of one cell to form two new cells. Usually the term is applied to the production of a new individual (either asexually, from a single parent organism, or sexually, from at least two differing parent organisms), although strictly speaking it also describes the production of new cells in the process of growth.

Looking at this list, the feature most difficult to mimic with engineering is reproduction. Life is particularly good at this, and so perhaps this remains the best dividing line between the living and the non-living.

A long discussion offering a variety of alternate definitions of life in the Encyclopedia Britannica concludes by pointing out the particular problem biology currently faces; we only know about one form of life, that on Earth, and all of the life on Earth seems completely related.

“The existence of diverse definitions of life surely means that life is something complicated. A fundamental understanding of biological systems has existed since the second half of the 19th century. But the number and diversity of definitions suggest something else as well. As detailed below, all the organisms on the Earth are extremely closely related, despite superficial differences. The fundamental ground pattern, both in form and in matter, of all life on Earth is essentially identical. As will emerge below, this identity probably implies that all organisms on Earth are evolved from a single instance of the origin of life. It is difficult to generalize from a single example, and in this respect the biologist is fundamentally handicapped as compared, say, to the chemist or physicist or geologist or meteorologist, who now can study aspects of his discipline beyond the Earth. If there is truly only one sort of life on Earth, then perspective is lacking in the most fundamental way.”⁵

Among all this confusion there is a clear consensus that, whatever life is exactly, all living things exist in open thermodynamic circumstances in which they take in resources, use them to construct themselves and

⁵ **life.** (2007). In *Encyclopædia Britannica*. Retrieved August 25, 2007, from Encyclopædia Britannica Online: <http://search.eb.com/eb/article-9106478>

near replicas of themselves, then expend these resources, always at the expense of substantial increases in entropy.

We have seen that outcomes which increase entropy are statistically likely; so likely as to be inevitable. In this way of thinking, life may be made likely by the same mechanisms which cause other cyclic phenomena in open thermodynamic systems like weather on the Earth. If this is so, life (whatever it is) should exist anywhere in the universe where conditions allow it. This prediction presents science with one of its most tantalizing challenges for the future. If we understand life, we really should find it in many places. There's a real chance this will happen during your life, and perhaps one of you will find it.

Life and motive force

Another feature we identify very strongly with life is the ability to get up and move around. Granted, much of life is vegetative; moving little, if at all. Still, for most of human history, the ability to move unaided by something like gravity or the wind was the sole province of life. The only way to reliably get things done, to act in the world, was to get a living thing to do it. People have always applied their own motive force to transport themselves and do things like digging ditches and harvesting grain. Very early on, they began to appropriate the motive force of other animals; riding horses, yoking oxen, using dogs and falcons to chase down prey. Borrowing the motive force of animals helped to enable a first great round of human population and cultural growth. We still recognize this heritage when we measure the power output of a system: 1 horsepower is approximately 745 Watts.

There is a long history of attempts to convert thermal energy to motive force through engines. The earliest known is the so-called aeolipile of the ancient world, described by Hero of Alexandria around the year 50 AD, among others. This device is really a kind of rocket, in which steam produced by boiling water, escaping through a pair of oppositely directed nozzles, generates a torque which rotates the vessel. It is not known whether this device was used to accomplish anything other than to delight. It seems to have been used primarily to impress, and it was wonderful; a device which could turn heat into organized motion. It seemed alive.



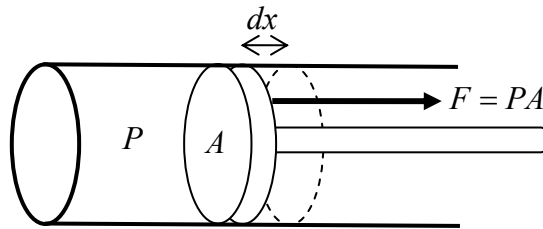
It took more than 1500 years before a new generation of steam engines would be harnessed to provide access to seemingly limitless motive force. Steam engines could apply larger forces, at higher speeds, than draft animals. They didn't tire, and could make use of seemingly limitless supplies of coal. The first widely known and effective commercial engine was the Newcomen engine, used to pump water from mines. This was followed by a flurry of innovation, especially in the hands of James Watt and his collaborators. These early inventors made their progress with no deep scientific understanding, tuning their devices with great skill, but lacking a theory to guide their understanding of efficiency or to tell them what limits might apply.

During the early 19th century a group of European scientists, including Sadi Carnot in France, William Rankine and William Thompson (later Lord Kelvin) in Scotland, and Rudolf Clausius and Hermann Helmholtz in Germany, developed a rich and powerful theory called thermodynamics which explains the action of engines, provides guidance on how to optimize them, and shows clearly what their fundamental limits are. Along the way, they understood the conservation of energy, invented the concept of entropy, and were for the first time able to explain why some things happen in the world while others don't.

16.4 Thermodynamic cycles and engines

One way of better understanding how the complex engines of life might work is to consider simple, man-made thermodynamic cycles and the mechanical engines which they drive. Let's define one first. A thermodynamic cycle is a system which starts in some state, then goes through a series of changes, returning eventually to the original state, while doing something along the way. This might be an engine going through a single power stroke, consuming some fuel along the way, then returning to its initial state. Or it might be a muscle cell, suddenly tugging itself shorter, using up some ATP along the way, then returning to its initial state.

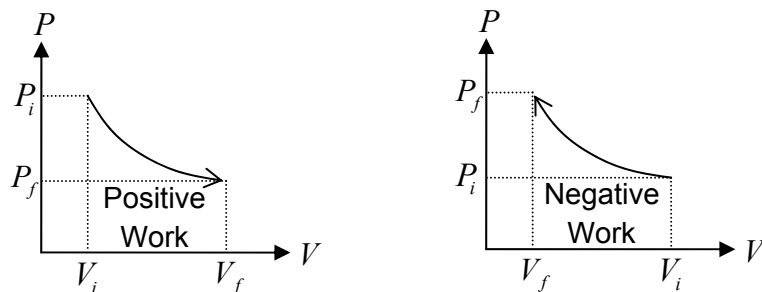
To explore such thermodynamic cycles, we'll consider a simple system: some quantity of ideal gas contained within a piston which begins at some initial temperature T_i , pressure P_i , and volume V_i . As it goes through a cycle, T , P , and V will change in some way, eventually returning to the initial state. This ideal gas model system is useful to discuss partly because it is simple, but also because it actually underlies many of the engines humans use to create lifelike motive force. There are a semi-infinite variety of other possibilities we could consider, and it is important to remember that this little enclosed sample of ideal gas is only one of many.



Imagine you allow such a gas to expand a bit by pushing the piston outward through some distance dx . Since the gas is pushing on the piston with a force $F = PA$ (with A the area of the piston), the gas must do work on the piston:

$$W_{\text{gas on piston}} = Fdx = PAdx = PdV$$

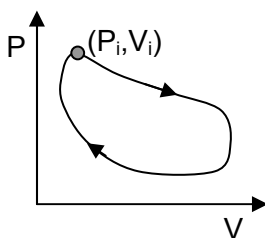
In the last step we noted that $A dx$ is just the change in volume of the gas dV . As we move the piston out, the pressure may change too. But we can still track the total work done by the gas by adding up the quantity PdV in each little step dx .



This is clarified by considering the whole ‘path’ of the system in the P, V plane. It starts at some point (P_i, V_i) , and moves step by step to a new position (P_f, V_f) . For each little change in volume dV , the gas does work PdV . So the total work done by the gas is the area under this (P, V) curve. Early investigators of steam engines called this the ‘indicator diagram’.

Any time the gas expands, the change in volume dV is positive, and the gas does positive work. When the gas is doing positive work, it is transferring some of its internal energy to whatever the piston is connected to. But of course it’s possible that something is putting energy back into the gas (perhaps as heat); all we’ve done here is to insure that the gas is doing work by pushing on the piston. Any time the gas is compressed, the change in volume dV is negative, and the gas does negative work. In this case, energy is being transferred into the gas. Of course it’s possible that something adds to or removes energy from the gas by some means other than work, perhaps as heat. What we see here is how we can track the changes which occur in the gas by examining how it moves through the P and V state space.

In a thermodynamic cycle, P and V start out somewhere, then change continuously, eventually returning to their starting point. The figure below shows an example. To understand such thermodynamic cycles in detail, it is helpful to consider a few particular kinds of processes which the gas might undergo. We will then use these particular processes to construct thermodynamic cycles for use in machines.



Isothermal processes

If we change pressure and volume together in just the right way, we can construct a process in which the temperature of the gas stays the same: an isothermal process. Any time a gas expands or contracts, it is doing work; either gaining or losing internal energy. To keep its temperature the same, something must

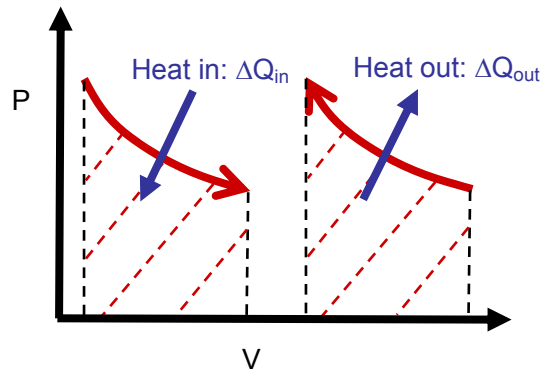
either remove or add some heat while the gas expands or contracts. This might be accomplished by placing the gas in contact with a large heat reservoir. Then, if the gas gets a little cooler due to expansion, heat can flow in, replacing the energy lost due to expansion. If it gets hotter due to compression, heat can flow out, removing the energy gained in compression.

When the process is isothermal we can use the ideal gas law to tell us exactly what the path is in the (P, V) plane.

$$PV = nRT \quad \text{implies} \quad P = \frac{nRT}{V}$$

In words, the pressure is inversely proportional to the volume.

This figure shows both an isothermal expansion (which requires putting heat in to maintain the temperature) and an isothermal contraction (which requires removing heat to maintain the temperature). This particular path in the plane, with $P \propto 1/V$, is the one for which T remains the same. Any path on which P falls more rapidly than this is one in which the gas temperature drops. Any path where P falls more slowly than as $1/V$ is one in which the gas temperature rises.



For example, an “isochoric” path which maintains constant volume but changes pressure would be a line straight up or down in this plane. If you go straight down, the pressure is falling more steeply than isothermal, and the temperature must be falling. If you go straight up, the pressure is falling more slowly than isothermal (in fact it’s increasing) and the temperature must be rising.

How much work is done in this isothermal expansion? We know that the work is given by the definition:

$$W = \int_{V_i}^{V_f} PdV = \int_{V_i}^{V_f} \left(\frac{nRT}{V} \right) dV = nRT \int_{V_i}^{V_f} \left(\frac{dV}{V} \right) = nRT \left[\ln(V_f) - \ln(V_i) \right]$$

$$W = nRT \ln \left(\frac{V_f}{V_i} \right)$$

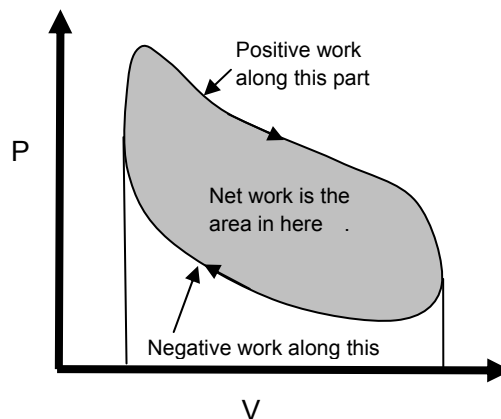
This is the amount of work done by the gas during an isothermal transformation. Notice that it depends on both the temperature at which it occurs (it would do more work if the temperature were higher) and on the change in volume V_f/V_i (it would do more work if the change in volume is large).

How much heat has to be put in to make this isothermal process happen? The amount of heat put into the gas has to exactly balance the energy the gas loses by doing work, or else the temperature will change, so:

$$Q_{\text{in}} = W_{\text{out}} = nRT \ln\left(\frac{V_f}{V_i}\right)$$

Constructing a cycle, the Stirling Cycle

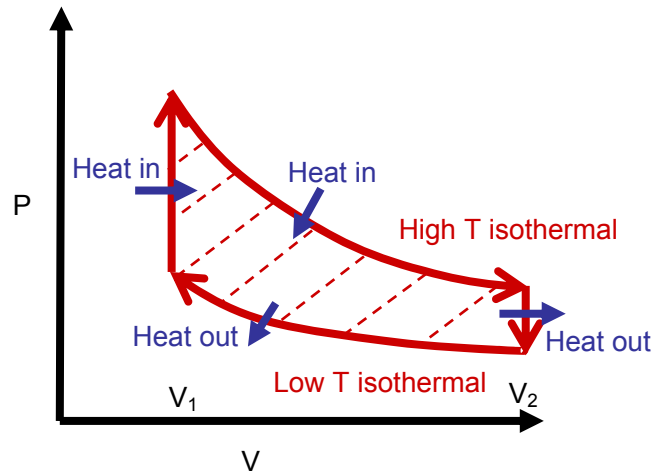
Any closed path in this (P, V) plane would be a thermodynamic cycle, starting, moving around doing some work etc., and then returning to the same initial condition. How much work would be done by a cycle? If the cycle goes clockwise, as shown, positive work is done during the expansion on the top, while negative work is done during the contraction on the bottom. The positive work is the area under the top curve, while the negative work is the area under the bottom, so the net work is the area in between the two.



To complete the cycle, we'd have to put a net amount of heat into the cycle which is just equal to the total work done. Otherwise we wouldn't be able to return the gas to exactly the same point in the (P, V) plane it started at. From this, you can start to see what a thermodynamic cycle like this does. To complete it, you put thermal energy (heat) in, and you get mechanical energy (work, pushing on the piston) out.

That's what engines like steam engines do, and they are central to technology because they allowed people, for the first time, to harness thermal energy to accomplish mechanic work, to create the kind of motive force previously available only from life. For the first time, people were able to do work without using a living thing. Tasks like moving things around, grinding wheat into flour, or operating machines, which previously required the labor of a living thing, could now be done by converting some of the energy present in heat and turning it into motive force.

To get a better sense of what such a thermodynamic cycle might be like, we will consider in detail one particular cycle, though there are in fact infinitely many which are possible. This one is called the “Stirling cycle” because it was invented by Robert Stirling around 1816. The Stirling cycle engine is sometimes called an ‘air engine’, because the gas which expands and contracts in the cycle is just air. Earlier steam engines had allowed hot steam to expand, afterwards sweeping out the cool gas and introducing new steam each time. In the Stirling cycle you take the gas inside (which could be air, or any other gas which behaves reasonably like an ideal gas) through a series of four steps which are illustrated in the (P, V) diagram below.



1: Start out with an isothermal expansion at some high temperature T_H from volume V_1 to V_2 . During this phase, the gas does positive work, pushing the piston outward. You could use this part of the work to do something. The amount of positive work done is given by

$$W_H = nRT_H \ln\left(\frac{V_2}{V_1}\right)$$

To maintain the temperature during this expansion you have to somehow add heat to balance the energy lost when the gas does work. The amount of heat you add is just equal to the work done: $Q_H^{\text{in}} = W_H$.

2: After this expansion phase, the gas in the engine undergoes an isochoric (constant volume) decrease in pressure. This is also a decrease in temperature to a new temperature T_C . You can see that the temperature must decrease in an isochoric decrease in pressure by recalling the ideal gas equation

$$PV = nRT$$

And imagining what happens if the pressure decreases while the volume is held constant. Since the temperature decreases in this process, heat must be removed from the air in this step. How much heat is removed? This depends on both the molar heat capacity of the material and on how much gas is in the piston:

$$\Delta Q_{\text{right}}^{\text{out}} = nC_{\text{gas}}(T_H - T_C)$$

3: Now the gas undergoes a new isothermal compression from volume V_2 back down to the original volume V_1 . During this compression, the gas does negative work (that is, work is done on the gas). To keep the temperature of the gas constant at a constant temperature T_C while compressing it, heat must be removed. The amount of heat removed must just balance the amount of work going into the gas

$$W_C = nRT_C \ln\left(\frac{V_1}{V_2}\right)$$

4: In the final step, you add back heat to take the gas through an isochoric process back up to the hot temperature T_H . This requires you to add heat:

$$\Delta Q_{\text{left}}^{\text{in}} = nC_{\text{gas}}(T_H - T_C)$$

Note that in this step you're just putting back exactly the same amount of heat you took out in step two.

What's the total work done each time the gas goes through this cycle?

$$W_{\text{total}} = W_H - W_C = nR(T_H - T_C) \ln\left(\frac{V_2}{V_1}\right)$$

The amount of work done in each cycle could be increased by either making the volume change larger, or by making the temperature difference between the hot and cold ends larger.

Efficiency in thermodynamic cycles

If you think of a thermodynamic cycle like this as something which converts random thermal motion into work (something like ordered motion) it's worth asking how efficiently it does this. One way to measure this would be to compare the amount of work done to the amount of heat put in. For this Stirling cycle, the heat put in is ΔQ_H . You don't have to worry about the heat put in on the left ($\Delta Q_{\text{left}}^{\text{in}}$), because it is perfectly balanced by the heat taken out on the right ($\Delta Q_{\text{right}}^{\text{out}}$). If we compare the heat put in to the work out, we find:

$$\text{Efficiency} = \frac{\text{Work done}}{\text{Heat in}} = \frac{nR(T_H - T_C) \ln\left(\frac{V_2}{V_1}\right)}{nRT_H \ln\left(\frac{V_2}{V_1}\right)}$$

$$\text{Efficiency} = \frac{(T_H - T_C)}{T_H} = 1 - \frac{T_C}{T_H}$$

What does this mean? It says that, unless $T_C = 0$, the efficiency of this kind of cycle cannot be 100%. But it offers the possibility that, with T_H very large and T_C very low, a mechanism like this might allow you to turn *most* of the heat you put in into work. Remember too that this is the theoretical limit. Anything which happens during the cycle which creates any additional loss, like losses due to friction, or not being able to completely recycle the heat you took out on the right and put it back in on the left, will lower the efficiency still further.

Let's think about this practically though. Here on Earth, there are limits to how hot the hot end can be and how cold the cold end can be. Generally speaking, the "cold" end will be at the ambient surrounding temperature, while the hot end will be something like the temperature of boiling water. So we might have $T_H = 373 \text{ K}$ and $T_C = 300 \text{ K}$. This would give an efficiency of:

$$\text{Efficiency} = 1 - \frac{300 \text{ K}}{373 \text{ K}} = 0.2$$

That's the very best you can do with those constraints. Only 20% of the heat you take from the boiling water on the hot side of the cycle will get turned into work by this Sterling cycle engine. What happens to the rest? The rest of the heat gets pulled out of the hot reservoir and dumped into the cold reservoir. That is, *it gets spread out* throughout the part of the world which is at the ambient temperature around 300 K.

Why not just run the cycle again? Take that heat you remove back up to 373 K and send it through again. The problem is, you can't do that. Once it's spread out at the low temperature, you can't get it back up to the high temperature. So the heat you expel at low temperature, 80% of what you took out the boiling water, is wasted, lost in an unrecoverable way.

This is the key lesson of thermodynamic cycles. They can create order, taking random motion and converting *some* of it into ordered motion, but they can only do this at the cost of more effectively spreading out some of the heat taken in. Just like other thermodynamic processes which 'just happen', life does this too. It gets things done, creates order, moves around, reproduces. But all this happens only at the cost of speeding the collective increase in entropy.

16.5 How life works: 'Gibbs free energy' and life

The thermodynamic engines so widely used in our technology powered the industrial revolution. They provided our first real insight into topics like entropy, and helped us to understand why some things happen, other things don't, and even in a world where energy is conserved, efficiency is always less than perfect. But these heat engines, lively though they are, function in a manner very different from life. They extract work from heat through expansion and contraction at different temperatures. Life, by contrast, does what it does in conditions which essentially never change temperature, pressure, or volume. How then does life work?

A system will reach the equilibrium of maximum entropy when all the energy in it is spread equally among all the forms it might take. Once this happens, the system will never change. To function continuously, it is not enough for life to contain energy; it must have energy flowing through it. That energy must enter in a form not yet fully spread. In the process of spreading, it can accomplish something;

create some local order, do some work. By releasing all the energy previously trapped in some not-yet-spread form, life actually aids the increase in entropy of the universe.

Energy in this kind of not-yet-spread form is called ‘Gibbs energy’, or less formally, free energy. It can be extracted and used for some purpose. Energy which is already fully spread is not free energy, and nothing can really be done with it. For a system at fixed temperature, pressure, and volume (a good approximation for life), the Gibbs energy can be written:

$$G = U - TS$$

In this equation, G is the Gibbs energy, U is the total internal energy in the system, T is the temperature, and S is the entropy of the system.

In a living system, energy enters (U is increased) in a form not yet fully spread. For life on Earth, this almost always begins with the absorption of some sunlight by a plant. Photosynthesis in the plant can capture this incoming energy, allow some of it to escape (increasing the entropy somewhat) and store the rest in a form (usually sugars) from which it can be extracted later. The energy now in the sugars is still not fully spread. Animals may eat these plants, release the free energy they have stored, and hence move toward still more complete spreading energy which ultimately derives from the Sun.

Thermodynamic engines do something similar. They take in energy at high temperatures, from a place where the energy is very concentrated, do work, then expel the remaining energy at lower temperatures, allowing it to spread substantially. In fact, changes in entropy were originally quantified using this definition:

$$dS = \frac{dQ}{T}$$

So that the absorption of heat (dQ) at high temperature would correspond to a smaller change in entropy than the expulsion of the same amount of heat at lower temperature. An engine executing the Stirling cycle would take in energy with a small amount of entropy on the high temperature side, then expel that energy with a lot of entropy on the low temperature side.

Living systems do something very similar, except that the energy enters not as heat, and the whole process takes place at (essentially) fixed temperature. But once again, the energy arrives with little entropy and leaves with more. Given the definition of Gibbs energy above, you can see that increasing entropy corresponds to decreasing free energy. Life lives off of the free energy arriving from the Sun. Once we’re done with it, the free energy is substantially reduced. Living things help along the inevitable spread of energy and increase of entropy. Like thermodynamic engines, living things can do what they do because they aid the spread of energy.

Statistical physics tells us that the increase of entropy in large systems is so likely as to be inevitable. Perhaps this makes life, an effective mechanism for increasing entropy, so likely as to be inevitable. If this is so, life should emerge quite freely, and exist pretty nearly anywhere it can.

A Quick Summary of Some Important Relations

Formation of structure while increasing entropy:

In a closed system, order will not emerge spontaneously; this would require a decrease of entropy. In an open system, energy can spread, increasing entropy while leaving behind matter in a more ordered form. The net entropy increases, even when locally entropy is decreasing. Snowflakes and spherical planets provide nice examples.

Definitions of life:

Defining life precisely has proved impossible. The ability to use environmental resources to self-replicate seems the most important dividing line.

Thermodynamic cycles and motive force:

Engines which can replace the motive force provided by animals powered the industrial revolution. They work by taking in heat, converting some of it to mechanical work, and expelling the rest at low temperature.

PV diagrams:

When gas changes pressure and volume, it may change temperature and/or do work. In this process, its internal energy may change.

$$W_{\text{done by gas}} = \text{area under PV curve} = \int_{V_i}^{V_f} PdV$$
$$\Delta U_{\text{internal}} = nc_{\text{molar}}^{\text{constant } V} \Delta T = \Delta Q_{\text{in}} - W_{\text{done by gas}}$$

For an ideal gas, knowing the pressure and volume determines the temperature, so you can always find the change in temperature for any process by using the ideal gas law at the beginning and end.

Special processes:

In isothermal processes, the work done by the gas is replaced by heat which flows in so that the temperature of the gas remains the same

$$W_{\text{done by gas}} = nRT \ln \left(\frac{V_f}{V_i} \right) = Q_{\text{in}}$$

In an isochoric process, the volume remains unchanged, no work is done by the gas, and heat flows in or out to change the temperature. In an isobaric process, the pressure remains unchanged, work is done by the gas, and heat flows in or out as well.

17. Floating: fluid statics, including surface effects

- 1) What is a fluid and how can you tell?
 - i. Fluids and solids
 - ii. Liquids and gases
- 2) Hydrostatics
 - i. Gravity and fluid pressure
 - ii. Pressure in water
 - iii. Atmospheric pressure: gravity and a compressible fluid
 - iv. Pressure measurement and Pascal's principle
- 3) Hydrostatic pressure and buoyancy
 - i. Archimedes principle
 - ii. Floating, sinking, and the buoyant force
 - iii. Some consequences for life in water
- 4) Liquids have surfaces: life at the interface
 - i. Surface energy
 - ii. Surface tension
- 5) Consequences of surface tension
 - i. Walking on water: an array of forms
 - ii. Bubbles and drops
 - iii. Soap films, bubbles, and membranes
- 6) Solid-liquid-gas boundaries: wetting and not
 - i. The contact angle
 - ii. Non-wetting surfaces
 - iii. Capillary flow
 - iv. Surface physics for solids: importance for nano-technology

Physics for the Life Sciences Chapter 17

17.1 A new subject: fluids

The Earth is a wet, airy place, distinguished from other known planets by the presence of a warm atmosphere and surface liquid water. All the life we know exists in, and is largely made of, two different fluids; air and water. Being fluids, the two share many properties. But they differ in essential ways as well. One is a gas, the other a liquid. We will see that this creates differences in both quantity (their densities differ by a factor of 1000) and quality (air is easily compressed, while water is as stiff as a solid).

Understanding the behavior of fluids will require us to deal with phenomena we haven't discussed before. When we introduced mechanics, we first treated things as point objects. Then we gradually worked out how to treat extended objects which can do things points can't; like rotate, or distort. Fluids do something further still; they flow. The ability of fluids to relatively freely alter the arrangement of their atoms creates new opportunities and challenges for life. Life moves within fluids, pushing against them to speed up, slow down, and generate lift. Living things pull fluids in, push them out, and use them in a form of forced convection to deliver nutrients, carry off wastes, and transfer heat.

To understand life more completely, we need to understand how fluids work. Doing this will provide a new appreciation for life on land, on the sea, and at the boundaries between them.

Fluids vs. Solids:

To start, let's consider what makes a fluid a fluid, and why this is different from a solid. We have already discussed how solids respond to the application of various static loads. We have seen that the basic response of a solid to a load is a deformation:

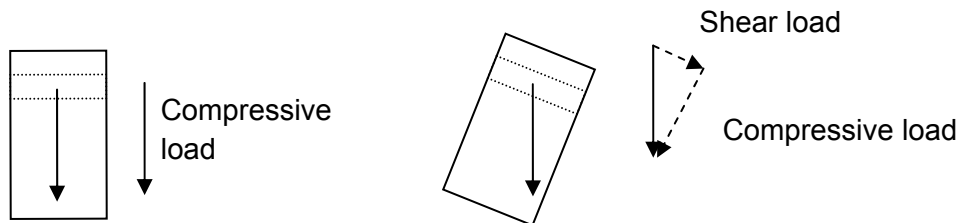
$$\begin{aligned}\sigma_{\text{tensile}} &= E\varepsilon_{\text{tensile}} && \text{for tensile or compressive loads} \\ \sigma_{\text{shear}} &= S\varepsilon_{\text{shear}} && \text{for shear loads} \\ \sigma_{\text{bulk}} &= B\varepsilon_{\text{bulk}} && \text{for hydrostatic loads}\end{aligned}$$

So although there are different kinds of loads, and different responses to them, the situation with solids is pretty simple. Solids respond to loads by deforming. Many solids, especially biological ones, don't behave in precisely this linear 'Hooke's Law' fashion. But they still respond to loads by deforming. Apply a stress to a material and it will undergo some deformation until the forces balance, and then it will *stop* moving.

Fluids exhibit a fundamentally different behavior when a stress is applied to them. Their response is so different it's somewhat difficult to imagine doing it. How do I apply a stress to a tank of water? You can't simply grab it and pull. So we'll start with a seemingly limited case. Consider a tank of water and imagine what happens if I apply a shear stress to it. We might do this in two ways.

First, I could 'grab' the top of a fluid and push it sideways. We might do this by placing something on top of the fluid, a board perhaps, and pushing it to the side. What would happen? Unlike a solid, the fluid doesn't just distort a bit and push back on me. Instead it actually begins to flow, moving in the direction of the force without limit.

There is of another way to apply a shear to a fluid like this; tilt it sideways and let gravity apply the shear. Picture a cup filled with water. When it is straight up and down, gravity applies only a compressive stress to a layer of the water. The water can stand this. Like a solid it does squash inward a bit, until the pressure in it balances the downward force of gravity. Tip the cup on its side, however, and the stress becomes a combination of compressive stress into the fluid and a shear stress along the surface. This is illustrated in the picture below.



You can see that in such a tilted fluid gravity exerts a shear on each layer. If I tilt a solid in this way, the same gravitational shear is applied. But the solid just deforms a little, bending to the side. If I tilt a cup of fluid like water in this way, instead of distorting, it immediately begins to *flow*. Indeed this flow starts so quickly and happens so fast that its hard to see, or really even to imagine, a tilted glass of fluid.

This is a key difference between a fluid and a solid. If you place a solid under a shear stress, it distorts by a fixed amount. If you place a fluid under a shear stress it **flows** instead of distorting. Schematically, this gets expressed as a different form for stress-strain equations.

For solids, we have the familiar:

$$\text{Shear stress} \propto \text{Shear strain}$$

For fluids we will have:

$$\text{Shear stress} \propto \text{Shear rate}$$

When a constant shear stress is applied to a fluid it leads to a constant shear rate, a fixed rate of flow.

How can a fluid do this when a solid can't? In a fluid, the amount of energy required to move one atom in the material past another is relatively small. In particular, it is smaller than the typical thermal kinetic energy which the atoms possess. Since they already have enough energy to move past one another, they do so with no special effort. Give them the slightest urging, a little bit of shear stress, and their motion past one another just happens.

Now consider a hydrostatic load. What happens when you fill a cylinder with a liquid, cap it tightly with a piston, and push down? The water in the cylinder is compressed a bit until it pushes back up just as hard as we push down on it. This is just like the response of a solid, and liquids at least are almost as 'stiff' as solids, with bulk moduli in the 10^9 N/m^2 range. It turns out that liquids also behave a lot like solids in hydrostatic tension too, though this is less commonly observed. We will talk about it in the context of siphons a bit later.

Gases in compression and tension are, as we've seen, different from liquids and solids. Exactly what happens when I compress a gas depends on whether I allow the temperature to change. The pressure the gas exerts back on me will generally change, but very slowly, approximately as the ideal gas law suggests:

$$PV = nRT$$

So if, for example, T is held constant and the volume is decreased the pressure in the gas will increase, but slowly. To double the pressure you have to halve the volume. Changing the volume of a gas is very easy compared to changing the volume of a liquid or solid. Think about how hard you would have to push to halve the volume of a brick.

We will treat both kinds of fluids (gases and liquids) in parallel most of the time, but it will occasionally be essential to remember the easy compressibility of gases.

Why do I have to put the fluid in a container before pushing down on it? After all, I can take a block of solid and just push down on it, applying a normal compression. If I try that with a column of fluid, it immediately escapes my pressure by squirting out the sides. Fluids don't support normal compressions and tension, but only 'hydrostatic' compression or tension which pushes in on every side equally. If you give the fluid any little avenue of escape, it will take it, immediately flowing out of whatever escape hole you provide.

Hydrostatic pressure like this is something which, in a sense, points in every direction. If the pressure at a point is P , the fluid will push away from that point with a force per unit area $F/A = P$ in *every* direction. Since it has no particular direction, hydrostatic pressure is a scalar quantity. Tell me how large the pressure is, and I know everything about it. In this sense it is quite different from the mechanical tensile or compressive stress, as each of these is a vector quantity, with both magnitude and direction.

17.2 Hydrostatics

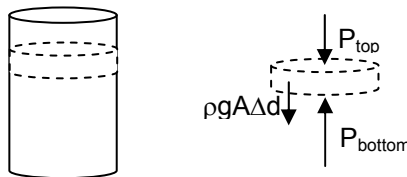
If you place a fluid in a cylinder and push down on the top with a tightly sealed piston (so that it can't flow out) you will generate a "hydrostatic" pressure inside it. Pressure is measured in force per unit area, and its typical units are N/m^2 . This unit, 1 N/m^2 , is also called a "Pascal".

What does hydrostatic pressure imply? If you push down on the top of a fluid with a certain force per unit area, that water will push outward, in every direction, on every part of the cylinder, with exactly the same pressure. This is because the fluid is doing everything it can to escape this pressure. In other words, if I push on the top with a pressure of 20 N/m^2 , the pressure with which the fluid will push out against the walls will increase *everywhere* in it by 20 N/m^2 .

Gravity and hydrostatic pressure

You're not the thing only which can create pressure in a fluid. Here on Earth at least, the fluid can do it on its own. Let's think about how the pressure in a fluid is affected by the weight of the fluid itself.

Consider a stable tank of fluid. Each layer is at rest, so the sum of forces acting on each layer must equal zero.



$$F_{\text{down}} = F_{\text{up}}$$

$$P_{\text{top}} A + \text{Weight} = P_{\text{bottom}} A$$

$$P_{\text{top}} A + \rho g A \Delta d = P_{\text{bottom}} A$$

$$P_{\text{bottom}} = P_{\text{top}} + \rho g \Delta d$$

Where I have used the fact that the mass of the disk is given by $m = \rho A \Delta d$. From this we can see that the pressure at the bottom of each layer of fluid is a little bigger than the pressure at the top. If we want to find the pressure at some depth below the top of the fluid, we just have to add together the increase in pressure which comes from each layer we pass through.

Consider a point a distance d from the surface. Adding up the pressure from each layer above this we find:

$$P_d = P_{\text{surface}} + \rho g d$$

The first term in this equation is the external pressure. Very often this is the pressure of the atmosphere; when the only thing pushing down on the surface is air. A typical value for this atmospheric pressure is about 10^5 N/m^2 , or 100 kiloPascals. Very often this pressure will be referred to as ‘one atmosphere’. This surface pressure on a fluid could be something different if, for example, we were actively pressing down on it, or had it enclosed in a container with high pressure gas above it.

The second term is the increase in pressure due to the weight of the fluid itself, which is sometimes called the gauge pressure. It's called the gauge pressure because often this is the pressure which we measure; the pressure *difference* between the outside (atmospheric pressure) and the pressure at some depth.

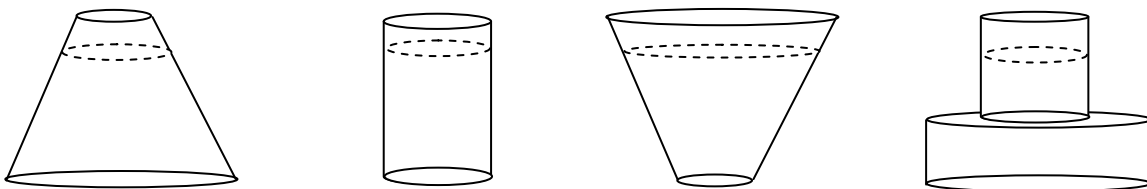
$$P_{\text{gauge}} = P_d - P_{\text{surface}} = \rho g d$$

For the two fluids we're concerned with, the densities are roughly $\rho_{\text{air}} = 1.2 \text{ kg/m}^3$ and

$\rho_{\text{water}} = 1000 \text{ kg/m}^3$. The large difference between these means that pressure will increase about a thousand times more rapidly with depth in water than in air. It's important to stress that gravity is the sole cause of these pressure increases. In the absence of gravity there would be no gauge pressure in the fluid, no increase in pressure with depth.

Why is hydrostatic fluid pressure different from what happens as the pressure increases within a stack of blocks? The pressure in the stack of blocks is directed, it acts only up and down, it does not act in every direction. This is why you don't have to push inward on the sides of a stack of blocks to keep them from squirting out sideways. With a fluid, you do have to do this.

Interestingly, the change of pressure with depth in a fluid is completely independent of the shape of the container it's in. Just to give an example, consider these four containers filled with water.



The change in pressure with depth in each of these four containers is exactly the same. This somewhat surprising fact is sometimes called the ‘hydrostatic paradox’, though of course it isn't really a paradox,

it's just a surprising fact. Perhaps it will convince you to think about how we could test this. Imagine connecting the three vessels at some point below the water surface with a pipe. Water could flow freely through this pipe. If the pressure at the chosen depth was higher in one container than another, water would flow from one to the other. This would raise the level of water in one container above that in the others.

Pressure in water

Because water is so dense, the pressure in it increases very rapidly with depth. For water, we can write:

$$P_{\text{gauge}} = \rho_{\text{water}}gd \approx (1000 \text{ kg/m}^3)(9.8 \text{ m/s}^2)d \approx (10^4 \text{ N/m}^2)d$$

Every time you descend in water, the pressure you feel increases by 10,000 N/m² or 10 kiloPascals for every meter deeper you go. That's roughly the weight of a 1000 kg object stacked on every square meter. The oceans have an average depth of about 3800 meters. At that depth, the pressure is 3.8x10⁷ N/m², or 380 atmospheres. In the deepest ocean trenches, the pressure rises a factor of three higher still, to more than 10⁸ N/m², or a thousand atmospheres.

The high pressures involved in living underwater (as most life does) have important physiological consequences. When you increase the pressure on a liquid, you enhance its ability to retain dissolved gases. This is why a closed can of soda retains dissolved carbon dioxide while it sits on the shelf; the pressure in the closed can is large. The extra pressure in the can is applied by the bit of extra gas sealed in the can above the drink; it's nothing to do with gauge pressure. Typically this extra pressure is a few atmospheres. When you crack the can open, you reduce the pressure on the soda, decreasing its ability to retain the dissolved carbon dioxide. The gas begins to emerge, forming the delightful fizziness which makes such drinks so charming.

Encountered in different circumstances, the same effect can be much less pleasant. Imagine that you dive a 100 meters deep. At this depth, the gauge pressure would be about 10⁶ N/m², or about 10 atmospheres, much higher than the pressure inside a can of soda. At this pressure, your blood can hold quite substantial quantities of dissolved inert gases, especially the nitrogen which will surely be present in the air you breathe. If you return to the surface too quickly, this dissolved gas will suddenly emerge from your blood, creating the frothy bubbles so pleasant in root beer. When they emerge in your blood vessels instead of your drink, however, you may die a painful death. The various consequences of pressure change make exploration of the deep ocean a technical challenge.

Atmospheric Pressure

Just as in any fluid, each layer of the Earth's atmosphere must support the weight of those above it. This creates a hydrostatic pressure which increases with depth from the vacuum of space to the Earth's surface. Everything on the surface of the Earth is submerged in an ocean of air. That air squeezes inward, pressing in on every exposed surface of every object with this atmospheric pressure.

At sea level (Ann Arbor is only 800 feet above this) the atmospheric pressure is about 10⁵ N/m². This is a very large pressure; every square meter of exposed surface has a force of 10⁵ N pushing on it. A 10⁵ N force is equal to the weight of ten metric tons. This may be hard to imagine, so consider a smaller area,

like the palm of your hand. This is maybe 8 cm x 8 cm, or 0.0064m². The force on this area, on the palm of your hand, is about 640 N. Hold your hand out palm up. Right now, there is a force pushing down on the palm of your hand about equal to the weight of a 65 kg person. How come you don't feel this person standing on your hand? While there is indeed a very large force pushing *down* on the top of your hand, there is an (almost exactly) equal force pushing *up* on the bottom of your hand.

Atmospheric pressure is hydrostatic; it surrounds you and pushes inward from every direction. It squeezes you inward, and if we were to remove it, you would expand outward a bit, with your atoms moving a little bit farther apart. How much would your volume change? We could estimate this by using what we know about the response of water to hydrostatic stress:

$$\sigma_{\text{bulk}} = B\varepsilon_{\text{bulk}} \quad \text{or} \quad \frac{F}{A} = B\left(\frac{\Delta V}{V}\right)$$

Putting in the change in pressure (10⁵ N/m²) and the bulk modulus of water (2.15x10⁹ N/m²) we find:

$$\frac{\Delta V}{V} = \frac{10^5 \text{ N/m}^2}{2.15 \times 10^9 \text{ N/m}^2} = 4.7 \times 10^{-5} = 0.0047\%$$

You'd expand a little, but hardly enough to notice.

Why do you feel outside pressure in a different way when you're under the water? For example, if you dive to the bottom of a pool, you feel a dramatic water pressure pushing in on your ears. Recall the difference between gases and liquids. Gases are easily compressed, liquids are not. So as you go underwater, the pressure of the water increases. This pressure increase is more than enough to compress the gas inside your eardrums. So your eardrum bends inward, stretching it in a painful way.

One more thing to note; if you have ever gone snorkeling, you may have wondered why snorkels are always so short, barely long enough to reach the air while you float on the surface. Imagine what would happen if you tried to a longer one which would allow you to (try to) breathe some distance d under the water. Outside you the water pressure would be: $P_{\text{atmosphere}} + \rho g d$, and inside your lungs, which remain directly connected to the air at the surface by the snorkel, the pressure would be $P_{\text{atmosphere}}$.

If you tried to do this at a depth of 10m, you would be squashed inward by the large pressure difference:

$$P_{\text{out}} - P_{\text{in}} = \rho_{\text{water}} g d = (1000 \text{ kg/m}^3)(9.8 \text{ m/s}^2)(10 \text{ m}) = 10^5 \text{ N/m}^2$$

This kind of pressure would quickly squash your chest inward, forcing out all the air. In practice, the greatest depth the typical person can handle with a snorkel is only a few feet. This is because you can expand your chest against a pressure only a small fraction of an atmospheric. If you try to do it deeper than this your chest will be compressed and all the air squeezed out. This will leave you with the feeling of having the wind knocked out of you.

How do divers go deeper? Snorkelers do it by holding their breath, closing off the connection between the air at the surface and the air in their lungs. Then when they dive underwater the air in their lungs can be

compressed without being forced out. As they return to the surface this air expands again, so that by the time they reconnect with the air through the snorkel it is again at normal atmospheric pressure.

Scuba divers use a fancier trick to allow them to go deeper still. The compressed air in their tanks is passed to their lungs through a “regulator” which matches the pressure of the air the diver receives to the pressure of the water which surrounds them. This generally works well, though there are important complications of breathing high pressure gas, especially the increased quantity of nitrogen which dissolves in the blood at high pressure.

What happens when the wind is knocked out of you? Usually you have air inside and outside your lungs at about the same pressure. So breathing in is just moving some air around. But if you *empty* your lung, then you have atmospheric pressure pushing in only, and it takes real effort to pull open your lungs again.

This idea also explains the operation of a ‘suction cup’. Are such things really “sucked onto” the surface to which they are attached? They seem to have remarkable properties, able to adhere strongly, then release freely, as if they could turn their attraction to the surface on and off. The answer lies with atmospheric pressure again.

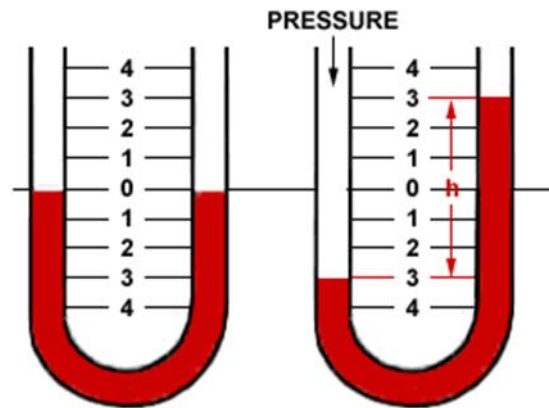
Imagine that I place an object on a surface and make it very flat, so that no air at all can be under it. Now the atmospheric pressure pushes down on the top, but there's no air under it to balance this, so it is pressed down very hard. That's what makes a suction cup work. It's not sucked onto the table, it is pressed in place by the atmospheric pressure. If you think about how you remove a suction cup this should be clear. You can't pull it straight off, that's much too hard. But if you peel up the edge just a bit, letting some air in underneath, it comes right off.

A drinking straw provides another example of a familiar atmospheric pressure driven device. We usually imagine that we are sucking our drink up the straw, but is this really what happens? When you take a drink, you close your lips around the straw, and then begin to expand your lungs. This slightly reduces the air pressure inside the straw. When this happens, the larger outside air pressure *pushes* the fluid up the straw.

It is worthwhile to stress that you can never “pull” something along using a gas. Gases don't support tension *at all*. A vacuum cleaner cannot suck things up, and you can't be “sucked” out of a punctured airplane or a spaceship. You can be pushed out by the high pressure air inside your plane, racing to expand into the low pressure area outside the hole. Once the air which was inside the plane is gone, nothing further happens, and so long as you have something to breathe, you can ride back to the surface in perfect safety.

Pressure Measurement

Pressure is most often measured in a “differential” way. A simple device for doing this is called a u-tube manometer (a word which just means gas measuring device...). When a liquid is placed in this tube it flows until the liquid level is the same on both sides. If you now connect one side to a chamber containing gas at higher than atmospheric pressure, the fluid is pushed down on that side, and up on the other, until added pressure from the column of fluid equals the difference in pressure between the chamber and the outside.



The difference in pressure between inside and outside the vessel is then given by:

$$P_{\text{in}} - P_{\text{out}} = \rho_{\text{liquid in tube}} gh$$

This difference in height h , which can be read off the scale, measures the difference in pressure between the atmosphere and inside. This is why the term ρgh is so often called the "gauge pressure"; it's the part of the pressure that you see on the gauge.

There are other kinds of pressure sensors, many of which are absolute (actually measuring the full pressure, rather than the pressure in excess of atmospheric pressure). Some of these are based on what are called "piezoelectric" materials. These are materials which generate a measurable voltage when you squeeze them. Measure the voltage and you get the pressure.

Pascal's Principle and the hydraulic press:

As we said earlier, if you squeeze down on the top of the fluid, that pressure is transmitted equally to every point in the fluid. Any change in external pressure is propagated throughout the fluid. This fact is widely used in a force magnifying device called a hydraulic press. If we apply a force on a small area piston connected to a fluid filled chamber we generate a pressure. This pressure is transmitted equally throughout the fluid. If, somewhere else, it encounters a large area piston, this same pressure will create a large force.

The mathematics of this is quite simple:

$$\frac{F_{\text{small piston}}}{A_{\text{small piston}}} = P_{\text{created}} = \frac{F_{\text{large piston}}}{A_{\text{large piston}}}$$

$$F_{\text{larger piston}} = F_{\text{small piston}} \left(\frac{A_{\text{large piston}}}{A_{\text{small piston}}} \right)$$

This kind of mechanism is another kind of "force magnifier", like the block and tackle, the inclined plane, and lever. The trick here, as in all the other systems which magnify forces, is that to raise the big side a distance, you have to push down on the small side much farther. Not surprisingly, the balance is given by:

$$d_{\text{large piston}} = d_{\text{small piston}} \left(\frac{A_{\text{small piston}}}{A_{\text{large piston}}} \right)$$

So the work done using Pascal's principle to lift with a small force is the same as it would be simply lifting with a large force. You have to push the small force through a large distance, rather than a large force through a small distance. You're just able to use a much smaller force to do the lifting than you otherwise would. Given that we are such force limited creatures, these mechanisms are very useful.



Hydraulic press: small cylinder circled.



Hydraulic brakes: small cylinder circled.

Hydraulic mechanisms like this are used in many machines which need to apply a really large force, but possess engines only capable of producing small ones. Two common examples are shown above. The first is a hydraulic press, used for manually applying large forces. The example shown is made for pressing molds in the manufacture of dental mouthpieces. The second example is a hydraulic brake. When a modest force is applied to the small cylinder the same pressure is transferred to the two large brake pads on either side, pressing them with great force against the inside of the wheel, and generating the large friction required to slow the car. In addition to these applications, hydraulic systems are used in a wide variety of construction equipment, things like bulldozers and cranes.

17.3 Archimedes principle and buoyancy

There is one very important consequence of the increase of pressure with depth in a fluid, used very extensively by life. We talked in the start of the class about how one of the greatest mechanical challenges for an organism is to support its own weight against the downward force of gravity. The increase of pressure with depth in a fluid can help with this.

If an organism is submerged in a fluid (we all are...) the fluid beneath the organism will always push upward on it with a force somewhat larger than the fluid above it pushes down. This difference, caused by the increase in fluid pressure with depth, generates a net upward force called buoyancy. As we'll see, these buoyant forces are small in air (where the pressure changes slowly with depth), but large in water. In fact, they are large enough in water to make resisting gravity a nearly non-existent problem for creatures that spend their lives submerged.

To figure out exactly how large this buoyant force will be, imagine a thin cube filled with water, submerged in the water. This could be made, perhaps, by sinking a very thin plastic cube filled with water. This cube would be stationary, in equilibrium, so we know that:

$$\begin{aligned}\sum F_{\text{vertical}} &= P_{\text{bottom}}A - P_{\text{top}}A - mg = 0 \\ (P_{\text{bottom}} - P_{\text{top}})A &= mg\end{aligned}$$

In words, this object experiences a net upward force caused by the pressure change which is just equal to the weight of the water in the cube. This upward force, which is generated by the variation in pressure as you go deeper in a fluid, is called the Buoyant Force.

$$F_{\text{buoyant}} = (P_{\text{bottom}} - P_{\text{top}})A = \text{weight of the water displaced}$$

Now imagine that you replace the water filled plastic cube with a cube of iron of exactly the same shape and size. This substitution won't change the pressure pushing on the block at all, as this is affected only by increasing with depth. So the buoyant force is in this case exactly the same. It might seem that it matters what's in this box. But this isn't right. The change in pressure in the fluid depends only on the distance below the surface (remember the hydrostatic paradox).

To recap the argument: we use the consideration of a cube of water to figure out how large the buoyant force is, and find that it's exactly equal to the weight of the water which is inside the object. Then we make the point that the size of this buoyant force is the same if I replace the cube of water with a cube of anything. The upward buoyant force on any object in a fluid is *always equal to the weight of the fluid which the object displaces*.

The shape of the object does not matter

We worked out the buoyant force by considering a cube submerged in the fluid. We did this because it was easy, with the pressure on the top and the pressure on the bottom taking on uniform values across the surfaces. We were also able to ignore the inward pressure on the sides. Using this, we got the simple answer that the buoyant force is equal to the weight of the fluid displaced.

But what's the buoyant force is the submerged object has a different shape? What if it's shaped like a fish, or a submarine, or a person? While this may not be as obvious, it is nonetheless true that the buoyant force is always just equal to the weight of the displaced fluid.

Floating, sinking, and the buoyant force

So now consider what happens to some object we submerge in the water. It is pressed upward with a force just equal to the weight of the water which it displaced.

$$F_{\text{buoyant}} = m_{\text{displaced}}g = \rho_{\text{water}}V_{\text{object}}g$$

so, the total force on it is:

$$\begin{aligned}\sum F_{\text{vertical}} &= F_{\text{buoyant}} - W = \rho_{\text{water}}V_{\text{object}}g - \rho_{\text{object}}V_{\text{object}}g \\ \sum F_{\text{vertical}} &= (\rho_{\text{water}} - \rho_{\text{object}})V_{\text{object}}g\end{aligned}$$

What does this mean? If $\rho_{\text{water}} < \rho_{\text{object}}$, the net force is down, and the object sinks deeper in the water. If $\rho_{\text{water}} > \rho_{\text{object}}$, the net force is positive, and the object rises toward the top.

What happens when it reaches the top? Part of the object goes above the surface, until the upward buoyant force balances the downward weight:

$$\rho_{\text{water}}V_{\text{in water}}g = \rho_{\text{object}}V_{\text{object}}g \quad \text{or} \quad \frac{\rho_{\text{object}}}{\rho_{\text{water}}} = \frac{V_{\text{in}}}{V_{\text{object}}}$$

In other words it floats with only a fraction of its total volume submerged. If its density is close to the density of water, it will float with most of its volume submerged. If its density is much less than the density of water, it will float with only a little of its volume submerged.

What happens if it sinks to the bottom? When it reaches the bottom, it is partly supported by the buoyant force, and partly by the bottom of the vessel. This is what happens with you when you stand in a swimming pool. Instead of the normally large normal force you feel between your feet and the floor on land, you feel a quite tiny little normal force between your feet and the bottom of the pool. The extra buoyant support you receive from the water supports most of your weight, and is part of what makes splashing about in the pool so delightful.

Some consequences of Archimedes principle and this idea

The impact of the buoyant force is completely different for organisms living in air and water. The reason is simple. Living things are made mostly of water, so to first order the weight of an organism is about $\rho_{\text{water}}V_{\text{organism}}g$. In the air, the buoyant force is $\rho_{\text{air}}V_{\text{organism}}g$, and since $\rho_{\text{water}} \approx 1000\rho_{\text{air}}$, the buoyant force experienced by air dwelling creatures is negligible compared to their weight.

People have learned how to make hot air and Helium balloons which are supported in air by the buoyant force, but no organisms use the buoyant force of air to substantially support themselves. Many organisms live in the air, but their support has more to do with fluid flow and fluid friction (which we will discuss in a bit) than with buoyancy.

In water, by contrast, the buoyant force on an organism is quite naturally almost equal to its weight. The net force on an organism submerged in water is:

$$\sum F_{\text{vertical}} = (\rho_{\text{water}} - \rho_{\text{organism}}) V_{\text{organism}} g$$

Since the density of most creatures is close to that of water, this net force is usually small, and may be either up or down. You probably know this from your own experience. When you swim, taking a deep breath (and hence decreasing your density) can make you positively buoyant, while expelling it (and increasing your density) can make you sink. Just that little adjustment is enough to change the balance.

The presence of this supportive buoyant force for organisms living submerged is the single largest difference between life on land and in the water. Things that live in water don't have to support their weight, while those on land do. This fact was one of the several serious challenges evolution had to overcome when creatures first began to move from the sea onto land.



This is also the principal reason why so many sea creatures seem utterly different from those on land. Enormous numbers of large invertebrates, supported completely by the buoyant force, exist there. Bring them on land and they become piles of goo, but back in the lovely, supportive ocean, they get along quite well. Many are aggressive, highly competitive carnivores, like the deep sea siphonophore shown in the picture.

The presence of the buoyant force in the water also eliminates the size restrictions which are so unavoidable on land. The blue whale, as far as we know the largest animal to ever live on Earth, can be so enormous only because it lives in the sea.

Buoyancy is also responsible for the phrase “the tip of the iceberg”. Unlike most materials, water has the property that ice, its solid form, is *less* dense than water, its liquid form. This has many important consequences, including some which have been emphasized in blockbuster disaster films. Approximate values for these densities of pure water are:

$$\rho_{\text{ice}} = 0.914 \times 10^3 \text{ kg/m}^3$$

$$\rho_{\text{water}} = 1.000 \times 10^3 \text{ kg/m}^3$$

so about 0.914/1.000 or 91% of a floating iceberg is below the surface. In the ocean, sea water has a slightly larger density than pure water, and the ice remains pure, so icebergs in the ocean float a little higher, with a bit more of their volume out of the water.

17.4 Liquids and their surfaces: life at the interface

Liquids are things which flow smoothly, like gases, but which hold together at their surfaces, like solids. They do not expand freely. So liquids and solids have something that a gas does not: a surface. In solids this is interesting, but not dynamic. Since atoms don't move around in a solid, surfaces never just change, they have to be altered by external influences.

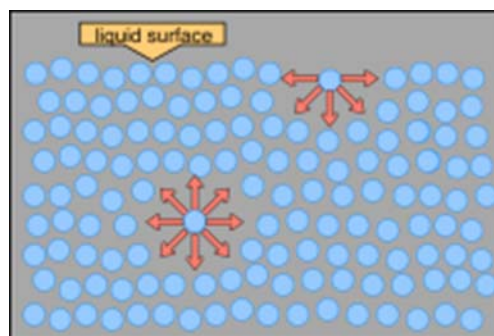
Liquids, by contrast, change their shape freely. They rearrange themselves on macroscopic scales, squirting away from anything that tries to shear them. On large scales, in rivers and big pipes, that's all you need to know. This is also about the only way large living things like you need to know about liquids.

When you look more closely and examine liquids on smaller scales a variety of really interesting things emerge. Small drops of fluid alter their own shape, pulling themselves together into drops as spherical as they can be. Drops of water on a waxed car "ball up" and roll off, while those on a warm frying pan spread out in a very thin layer. Water in a glass pulls itself up the edge a bit. These are all surface effects, and they become really important when you consider the behavior of fluids on small scales. As always, organisms living in this environment not only have to know about these properties, but actually take advantage of them in a wild variety of creative ways.

Surface energy

To understand the nature of this surface it is useful to consider what happens when an atom in a liquid moves around. Inside the liquid, the atom is attracted by its neighbors approximately equally in every direction, so it is relatively easy for it to move around. As it travels, it moves away from one atom, but towards another, so that on average it doesn't require much energy to move from place to place.

At the surface the situation is different. Here the atom only has neighbors behind it, and is only pulled back into the liquid. So when the atom attempts to move outward, it is pulled back into the liquid. This attraction, in a liquid, is large enough to keep atoms with typical thermal motions (kinetic energy $\sim k_B T$) from escaping. Atoms continually move to the surface, but each time they're pulled back, and remain in the liquid.



How is a gas different? In a gas the atoms are not connected to one another at all most of the time. An atom reaching the edge of a gas just keeps going and isn't pulled back by anything. So gases don't have surfaces and have no cohesion. This is also why you can't place a gas in tension at all, you can't stretch it.

And it is why a straw or a vacuum cleaner can't suck things up; it can only provide an empty space into which gas can push something.

What happens at the surface of a liquid is governed by random thermal motion, so to understand what happens we need to think about energy. As an atom moves from the bulk of the liquid to the surface, it moves against a restraining force, the same interatomic forces which hold the liquid together. Moving against this force means that work is being done. This work done by the rest of the liquid on the moving atom can be recovered if you let the atom turn around and go back into the liquid. So it makes sense to treat this "work done by the rest of the liquid" as a potential energy stored in the atom.

You can imagine an atom arriving at the surface with some thermal kinetic energy. It tries to keep going and exit the liquid, but the attraction of its neighbors pull it back inward, slowing it down, and stopping it (taking away its kinetic energy and replacing it with potential) and then reversing its motion, speeding it up, and sending it back into the liquid (reconverting that potential energy into kinetic energy).

What does this energy picture imply for the behavior of liquid surfaces? If I take a spherical drop and stretch it out into a thin sheet, I make its surface bigger, moving more atoms to the surface. This increases the amount of energy present in the form of surface potential energy. The more surface I have, the more energy in the liquid is stored in the form of this "surface potential energy". If I let this droplet go, that energy locked up in the surface will, as always, try to spread out. If the drop rearranges itself to be more spherical, excess surface energy can be released. Since this release of excess energy is so much more likely an outcome, it will happen freely. Other things being equal, *liquids will rearrange their shapes to minimize their surface areas.*

Here are some typical values of the surface energy per unit area for three liquids: (in contact with air)

Water	0.073 J/m ²
Ethyl Alcohol	0.022 J/m ²
Mercury	0.49 J/m ²

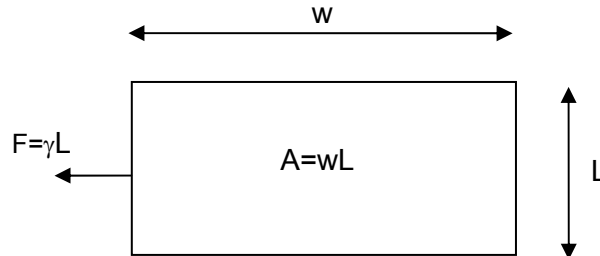
For example, it costs 0.073 Joules to create one square meter of surface area in water. That's not a whole lot of energy, which is why surface effects are not very important for large organisms. Notice too that these surface energies vary a lot. The value for mercury is about 7 times bigger than that for water. Mercury really sticks to itself.

Surface tension

There is an alternate way of thinking about what happens at the surface of a liquid, and these two equivalent descriptions are used interchangeably. To stretch the surface of a liquid (to add more atoms to the surface), we must apply a force *parallel* to the surface. This is because all I really have to do is pull two atoms at the surface apart, and another atom from the liquid below will move up to fill the space. It will be jostled there by ordinary thermal motion. The resistance of a surface to this kind of stretching is called the "surface tension" of the liquid, and it is measured in units of N/m, because the wider the surface is, the more force is needed to stretch it. The symbol γ is usually used to represent this.

Surface tension is the same thing as surface energy

Consider a little rectangle of the surface of a liquid. I have to apply a force $F=\gamma L$ in order to stretch the surface.



The *energy* associated with this whole rectangle is then the work done by this force in creating the surface:

$$W = \int \vec{F} \cdot d\vec{s} = Fw = \gamma Lw = \gamma A$$

where A is the surface area of the liquid. So the total surface energy of some area of a liquid surface is given the surface tension times the area.

We said a minute ago that the surface tension was expressed in terms of N/m. But $\text{N/m} = (\text{kgm/s}^2)/\text{m} = (\text{kgm}^2/\text{s}^2)/\text{m}^2 = \text{J/m}^2$. So the *same* number which represents the surface tension of a liquid is also the surface energy per unit surface area we discussed before. The number which represents the surface energy of water, 0.073 J/m^2 is also the surface tension 0.073 N/m .

Is the surface energy of water important to consider? The answer depends on scale, for the by now obvious reason that surface area varies like size^2 while volume varies like size^3 . When you have really large amounts of water, like an 8 oz glass, this surface energy is much smaller than other things like gravitational potential energy or bulk kinetic energy. For smaller amounts of water, like raindrops, surface energy and surface tension become much more important. So as we have often seen, the importance of this feature of liquids is scale dependent. We will see that surface phenomena play an extremely important role in a lot of small scale biology.

Some consequences of surface tension

Liquids are systems which have internal friction which can dissipate energy. Any extra energy which is in the liquid will, because of the inexorable increase in entropy, be given up and spread out into the surroundings. So if you start out with a liquid that has excess energy in some form, for instance a larger surface than it needs, it will move to make the surface smaller, and as it moves, dissipate the extra energy in the form of heat. This means that any fluid will, pretty rapidly, settle down to an arrangement in which the surface area of the fluid is minimized.

This makes systems which are governed by surface tension some of nature's most beautiful and elegant examples of minimization. We will look at a variety of different examples of how surface tension plays a role in the behavior of fluids, and you should keep in mind throughout this theme of optimization. Liquids change in ways which tend to spread out as much as possible their total energy. In part at least, this causes them to minimize their surface area. Surface tension is often what governs the shape of bodies of liquids.

Why do these liquids minimize their surface areas? What “forces” them to do so? Remember what we learned about statistical physics, about the idea of equipartition. Given a system with several ways of storing energy, you will find that random thermal rearrangement leads to the maximum sharing of energy among the different modes. How does this apply to a drop of liquid?

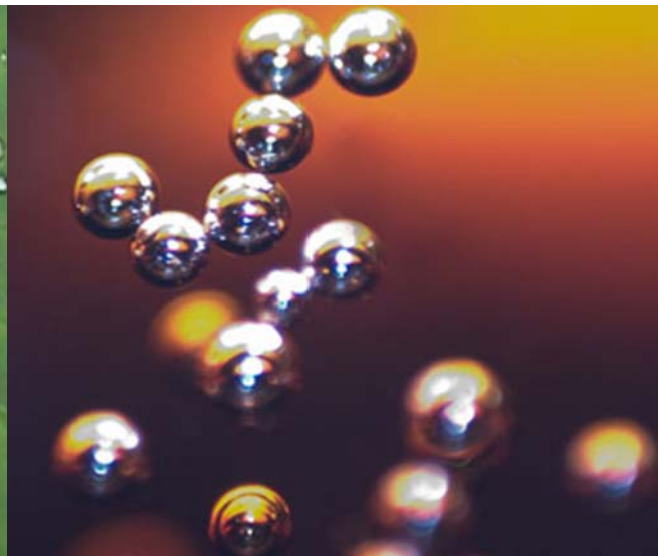
Imagine a drop of liquid is stretched out into a flat pancake. Such a drop would have a lot of energy in the form of this “surface energy”. If you could get this energy out of that form and spread it out into other forms like random thermal motion you would find your drop in a state with the energy more uniformly spread. The way to do this is to pull more of the atoms at the surface back into the liquid. When this happens, the bulk of the liquid will do *positive* work on the surface atom, reducing this surface potential energy and increasing the average kinetic energy of the atoms. So what happens is the drop pulls itself inward, doing its best to minimize its surface area.

This doesn't happen without limit however. If it did, every drop you see would be a perfect sphere. But as the liquid pulls inward to reduce the surface area, it must *raise* the average height of the liquid. If the drop is large, the increase in gravitational energy associated with this rise is larger than decrease in surface energy you get from shrinking it. In this case the equilibrium point will be an elliptical drop, with a surface area larger than a sphere, but with an average height lower than it would have as a sphere. In this picture of droplets on a leaf you can see that small droplets are much more spherical than large ones.

The same effect which leads water droplets to become spherical is also responsible for the spherical shape of air bubbles within water.



Water drops on a Lotus leaf



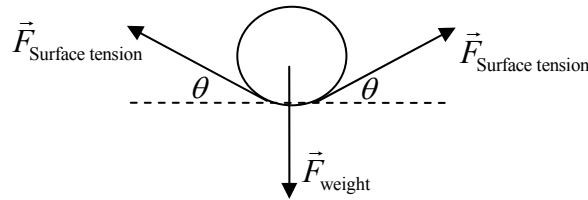
Tiny bubbles in a glass of cola

Surface tension can help to support something

Surface tension is able to apply an extra force to objects at the surfaces of liquids, often supporting them when buoyant forces alone could not. One famous example of this is a steel needle supported on the surface of water. The needle is much more dense than water, so it is not buoyancy which supports it. It really is the surface tension.



Based on this picture we can extract a free body diagram for the ‘floating needle’.



In this case the size of the surface tension force is fixed; it’s a property of water, and doesn’t care at all what kind of object we set on the surface. If this needle has a length L , then the total surface tension force on it is:

$$F_{\text{Surface tension}} = \gamma L$$

Like other tension forces, this force acts both ways along the surface, both left and right. If the needle has a very small weight, supporting it is easy, and the surface will be only slightly distorted; the angle θ will be quite small. If the needle is heavier, it will sink deeper into the water, making the angle θ larger. Summing the forces on this needle in the vertical direction, we can find an equation for the angle θ :

$$\begin{aligned}\sum F_{\text{vertical}} &= 2F_{\text{Surface tension}} \sin(\theta) - mg = 0 \\ \sin(\theta) &= \frac{mg}{2\gamma L}\end{aligned}$$

It seems remarkable that you can support a steel needle in this way. Obviously this won't work for a thick iron bar. What can we say about the limits? The mass of the bar can be written in terms of its density:

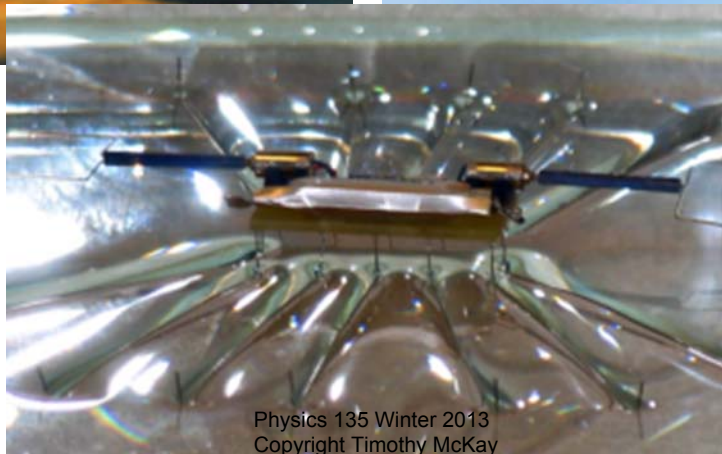
$$\sin(\theta) = \frac{\rho_{\text{needle}} (\pi r^2 L) g}{2\gamma L} = \frac{\rho_{\text{needle}} (\pi r^2) g}{2\gamma}$$

Anytime this quantity is larger than one, it will be impossible to find an angle for which surface tension will support the needle. This happens when:

$$\rho_{\text{needle}} (\pi r_{\text{needle}}^2) < \frac{2\gamma}{g}$$

Stated this way, it's a constraint on the linear density, the mass per unit length, of the needle. Putting in the numbers for water, we find that any needle with mass per unit length less than about 0.015 kg/m, or 15 grams per meter, can be supported in this way. This analysis suggests that surface tension could support a 5 centimeter long needle only if its mass is less than about 0.75 grams.

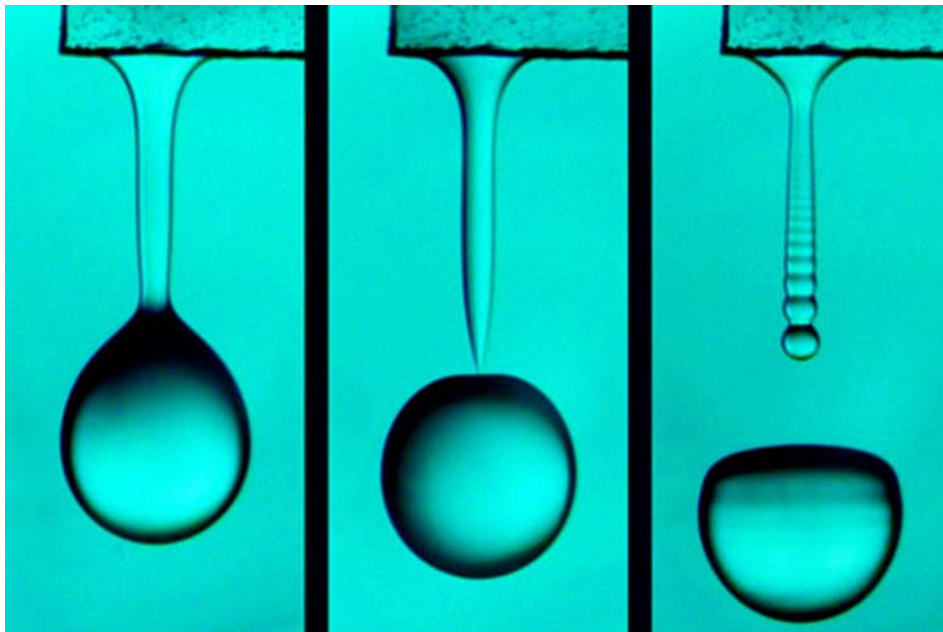
Remember though that there is also a buoyant force involved as well. As the needle pushes down into the fluid, it displaces water, creating an additional buoyant force. How large is this? If the needle displaces a volume almost equal to its own volume, this might be $\rho_{\text{water}} V_{\text{needle}} g$. This provides an additional supporting force.



Various organisms use surface tension, some in just this way. The most familiar is probably the water strider. Robot versions of water striders have been made as well.

In addition to needles, it's quite possible to "float" things on screens (like a window screen). In fact, a screen with sufficiently small holes will work just about as well as a solid object for keeping water out. This is why many fabrics, at least non-wetting ones (see below), can be waterproof even though they have many small holes in them. The leaf in the image on the previous page is also being supported largely by surface tension. You can see how the water's surface is bent downward around it.

Another example of the influence of surface tension on shape is the dripping faucet: The shape which a droplet forms is governed by gravity and surface tension. Water is held up in the pipe until the energy gained by lowering the gravitational potential energy of the fluid is enough to make up for the increased surface energy caused by stretching the drop out. As it starts to get long and thin, it becomes unstable, and finds a lower energy solution by pinching off a drop.



Frames drawn from a high speed video observation of a droplet forming and pinching off. These images come from a press release given out by the Taborek lab at the University of California at Irvine.

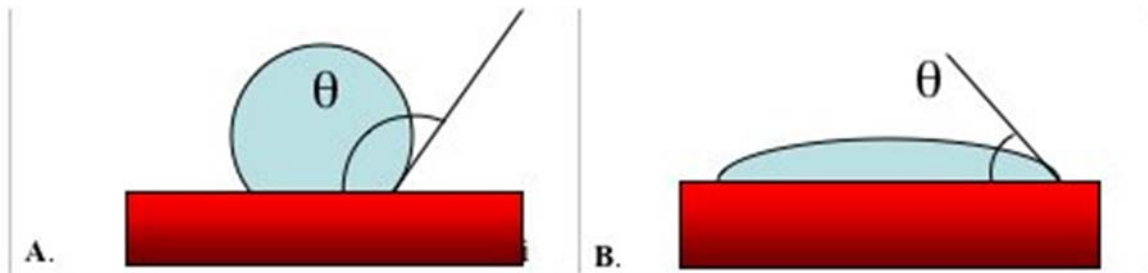
Interactions between liquids, solids, and gases

So far, the surface tension we've discussed is all about the interface between water and air. In this case, the surface energy is simple, and just related to the cohesion of the water. As a water molecule comes to the surface it is pulled back into the water, but never forward into the gas; the gas never pulls on the water molecules. When water is in contact with a solid (or indeed another liquid) the situation is more complex. Now it is necessary to consider the balance between cohesion (the forces which hold the liquid together)

and adhesion (the forces which attract atoms in the liquid to the solid with which it comes into contact). Since living things are made of solids and liquids, the interaction of water with them is very important.

One useful way to describe what happens when water encounters a solid surface as well as air is to keep track of is the “contact angle” between the liquid and solid in the presence of a gas.

If you put a drop of liquid on a surface, there will be some angle of contact at the point where the liquid first comes into contact with the solid. This “contact angle” is determined by a balance between the cohesion of the liquid and the adhesion between the liquid and the solid. If the atoms in the liquid are more attracted to other atoms in the liquid than to atoms in the solid the liquid will bunch together (as in diagram A). If the reverse is true, and the atoms in the liquid are more strongly attracted to the atoms in the solid than to their neighbors in the liquid, the liquid will “wet” the surface, spreading out over it as in diagram B.

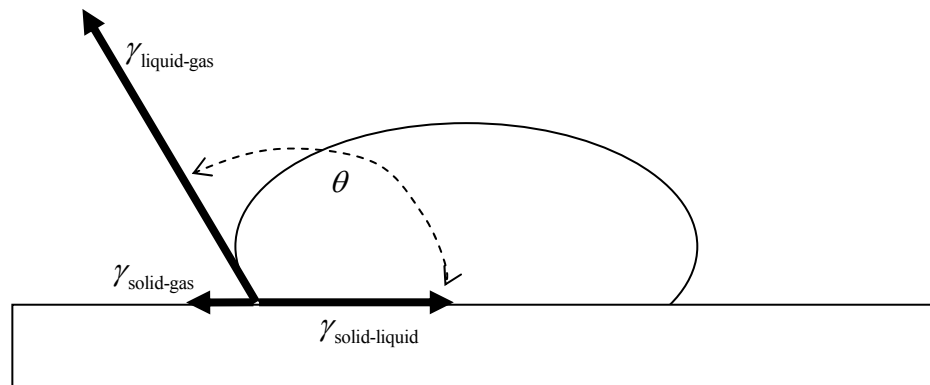


If the contact angle $\theta > 90^\circ$ the liquid does not "wet" the solid; the liquid is more attracted to itself than to the solid. If the contact angle $\theta < 90^\circ$ the liquid "wets" the solid; the liquid is more attracted to the solid than to itself. The value of this contact angle depends on the kind of liquid, the composition and state of the solid surface, and even to a certain extent on the gas which is around. It is worth noting, however, that there is no adhesion between the liquid and the gas.

The interaction between water and solids can vary a lot. Some solids are hydrophilic (water loving), and attract water strongly. These solids will tend to wet easily. Others are hydrophobic (water fearing). On these hydrophobic materials water tends to ball up into droplets and run off. You have probably seen this on well waxed cars. It is evidenced in the phrase “like water off a duck’s back”.

For the interface between a solid, a liquid, and a gas, the contact angle can be estimated from an understanding of the surface energies involved. As we have seen, the surface of a liquid tends to be minimized so that the energy contained in the surface can spread out, as energy typically does. Now we have to consider the surface energy of the liquid in contact with the gas $\gamma_{\text{liquid-gas}}$, the surface energy of the solid in contact with the gas $\gamma_{\text{solid-gas}}$, and the surface energy of the solid in contact with the liquid $\gamma_{\text{solid-liquid}}$. Recall that these surface energies also act like tensions, and when a drop on a solid surface

reaches equilibrium, these ‘tensions’ will balance. Modeling it in this way allows us to draw a diagram of these surface tensions:



In this picture, you can see that the balance of these tensions in the horizontal direction requires:

$$\gamma_{\text{solid-gas}} = \gamma_{\text{solid-liquid}} + \gamma_{\text{liquid-gas}} \cos(\theta)$$

$$\cos(\theta) = \frac{\gamma_{\text{solid-gas}} - \gamma_{\text{solid-liquid}}}{\gamma_{\text{liquid-gas}}}$$

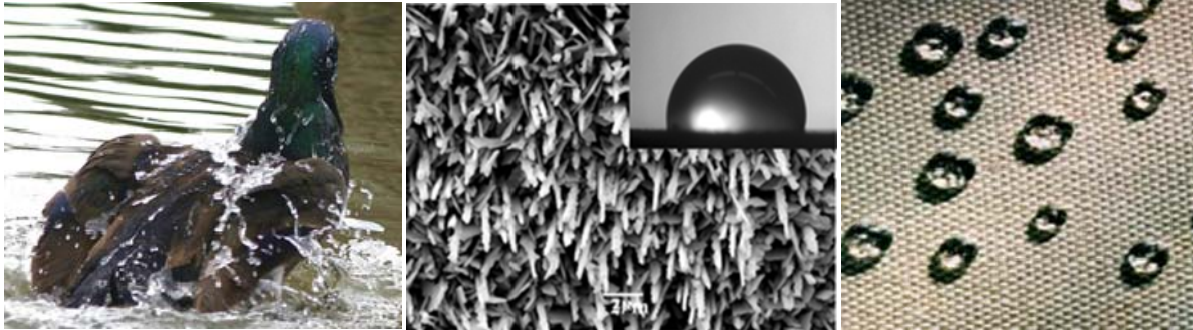
This is another relation originally worked out by Thomas Young around 1805.

When the surface energy associated with the solid-gas interface is larger than the surface energy associated with the liquid-gas interface, the cosine of the contact angle will be positive, and the contact angle itself will be less than 90° . When the surface energy associated with the solid-gas interface is smaller than the surface energy associated with the liquid-gas interface, the cosine of the contact angle will be negative, and the contact angle itself will be greater than 90° . Why is this? When the solid-gas surface energy is larger than the solid-liquid surface energy, energy can be released by spreading the drop across the surface; a spread out drop represents a lower energy state. When the solid-gas surface energy is smaller than the solid-liquid surface energy, it would cost energy to spread out the drop; a balled-up droplet like what's shown above is a lower energy state.

Staying dry and getting wet are two very important activities for living things. So it's not surprising that life has evolved some truly remarkable hydrophilic and hydrophobic materials. Most begin by altering the basic surface energy of the materials. When wetting is to be avoided, materials which are hydrophobic at the molecular level, such as oils and wax, are used. But there are many cases in living systems in which the effective surface energy is increased by changing the geometry of the surface. A common tactic for making seriously hydrophobic surfaces is to start with a chemically hydrophobic material. Remember this is a material which bonds less strongly to water than water does to itself. To make a water molecule move from the bulk of the water into contact with this material *costs energy*. Now to make it really cost a lot of energy, you might construct a material with a lot of surface. This is generally done by making the surface very rough, or even hairy. This way each little bit of area (viewed head on) actually has much more surface area than you would guess. This very large actual area makes the cost of wetting it become enormous.

Not surprisingly, this approach, which is widespread in living organisms, has been mimicked in human technology recently. So that now there are both hard surfaces and fabrics that shed water incredibly well.

Adhesion is a surface phenomenon, very complex, and under extensive study because of its technological importance. It can be extremely dependent on the detailed state of the surface (clean or contaminated for example). This is one of the main reasons water birds spend so much time preening. Keeping the feathers in good order is essential for shedding water, and shedding water is essential to avoid drowning or freezing to death.



Hydrophobic and aligned TiO₂ nanorod arrays

Liquid solid boundaries and capillary flow

The second important aspect of liquid-solid interfaces is capillary flow. When a fluid comes into contact with a vessel which it *likes* to adhere to, its contact angle will be less than 90°. This means that the adhesion of the liquid to the solid will stretch the fluid out, increasing its surface area. It can energetically afford to do this because the surface energy required to stretch the liquid is less than the energy gained by bonding more of the liquid to the solid.

Imagine what happens when you put such a liquid in contact with a *vertical* surface. For example, what happens when you place a hollow hydrophilic vertical cylinder down into the liquid? If you do this and the solid interacts well with the liquid, it will actually pull the liquid up the surface, suspending it from its surface tension.



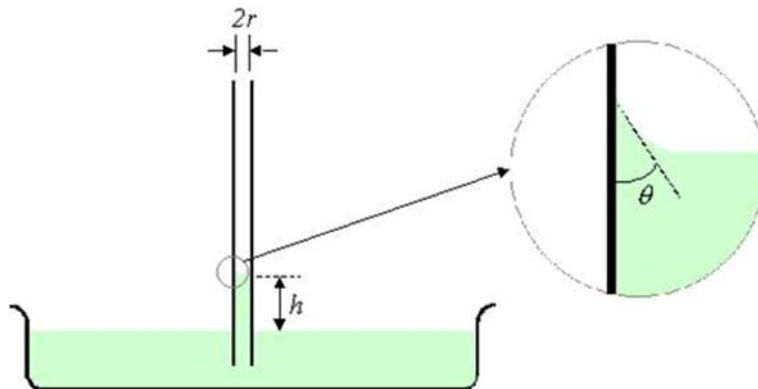
This process is enabled by random thermal motion. When you place the vertical surface in contact with the fluid, the tiny random motions in the fluid enable atoms near the solid to reach up the surface slightly.

When they do, they bond there, pulling the liquid behind them up as well. How rapidly this will happen depends on the temperature of the liquid as well as how strong the interaction between the liquid and the solid is.

How far can this adhesion pull a liquid up like this? Since the liquid is, in the end, hanging from its surface tension, the total upward component of the surface tension force is:

$$F_{\text{upward}} = \gamma_{\text{liquid-gas}} \times (\text{length of contact}) \times \cos(\theta)$$

Often, in large vessels, like a glass of water, this doesn't matter much. You can see the liquid pulled up at the edge, but it does little to the bulk. The situation is different when the tube you insert is very small. In this case you sometimes have enough force from this “capillary effect” to lift the fluid pretty far up the tube. Let’s work out how far.



The upward force for a circular tube like this is:

$$F_{\text{upward}} = \gamma_{\text{liquid-gas}} (2\pi r_{\text{column}}) \cos(\theta)$$

This force pulls up on a column of liquid with weight:

$$F_{\text{weight}} = \rho_{\text{liquid}} (\pi r_{\text{column}}^2 h) g$$

Capillary action will lift the fluid until the forces balance:

$$\rho_{\text{liquid}} (\pi r_{\text{column}}^2 h) g = \gamma_{\text{liquid-gas}} (2\pi r_{\text{column}}) \cos(\theta)$$

or

$$h = \frac{2\gamma_{\text{liquid-gas}} \cos(\theta)}{\rho_{\text{liquid}} g r}$$

In the limited case of perfect wetting, where $\theta = 0$, $h = 2\gamma_{\text{liquid-gas}} / \rho_{\text{liquid}} g r$.

How does this behave as I change the size of the tube? If the tube is big, like the size of a water glass, the liquid won't rise far at all. But if the tube is small, it can rise remarkably far. What sets the scale? The contact angle is determined by the fluid properties, the solid, and the conditions (temperature and pressure). The surface energy $\gamma_{\text{liquid-gas}}$ and density ρ_{liquid} are properties of the fluid.

Look at an illustrative example: a tube made of something which is nicely wet by water, so that the contact angle is zero. The liquid here is water, which has $\rho_{\text{water}} = 1000 \text{ kg/m}^3$ and $\gamma_{\text{water-air}} = 0.073 \text{ N/m}$. For this case we can write:

$$h = \frac{1.5 \times 10^{-5} \text{ m}^2}{r}$$

If the radius of this tube is 1 cm, water will rise in it by $1.5 \times 10^{-4} \text{ m}$, or 0.15 millimeters; not very much. But if the tube is thin, like 10^{-5} m or $10 \mu\text{m}$, then the water might rise up the tube a distance of 1.5 m! It uses thermal motion, combined with the natural attraction between the solid and the liquid, to pump itself up more than a meter.

Capillary action like this plays an important role in many parts of the life sciences, including both living systems themselves and in the technology used to study them. Perhaps most remarkable is the role it plays in the transport of water from ground level sources to leaves in the so-called vascular plants. These include pretty well all the large plants, from shrubs to redwoods. The roots of a tree are connected to its leaves by a continuous, narrow channel in the "xylem"; hollow, dead, tubes in the trunk which are very effectively wetted by water. These tubes have typical radii of $2 \times 10^{-5} \text{ m}$. For such a tube we would calculate a capillary height of

$$h = \frac{2(0.073 \text{ N/m})}{(1000 \text{ kg/m}^3)(9.8 \text{ m/s}^2)(2 \times 10^{-5} \text{ m})} = 0.8 \text{ m}$$

That's not very far, especially compared to the height of a tree. So how do trees manage to grow so tall? In fact this column of fluid is not being pulled up these relatively broad tubes. It is instead being held in place by capillary action in *much thinner* vessels called stoma which separate the cells in the leaves. These tiny stomata are incredibly numerous, and are responsible for moving CO_2 into the leaf and O_2 out. They typically have radii of $5 \times 10^{-9} \text{ m}$. How tall a column would this support? We have to be careful here, because the column is broad in the xylem, and only narrows at the top...

$$\begin{aligned} \text{Force available} &= 2\pi r_{\text{stoma}} \gamma_{\text{water}} \times N_{\text{stomata}} \\ \text{Force needed} &= \pi r_{\text{xylem}}^2 \rho_{\text{water}} g h \end{aligned}$$

Setting these equal, we find:

$$h = \frac{2r_{\text{stoma}} \gamma_{\text{water}} N_{\text{stomata}}}{r_{\text{xylem}}^2 \rho_{\text{water}} g}$$

What's the number of pores? We might guess that the total area of pores equals the total area of the xylem tube. This would imply:

$$N_{\text{stomata}} \times \pi r_{\text{stoma}}^2 = \pi r_{\text{xylem}}^2 \quad \text{or} \quad N_{\text{stomata}} = \left(\frac{r_{\text{xylem}}}{r_{\text{stoma}}} \right)^2 = 16 \text{ million}$$

Putting this into the above we would find:

$$h = \frac{2r_{\text{stoma}}\gamma_{\text{water}}N_{\text{stomata}}}{r_{\text{xylem}}^2\rho_{\text{water}}g} = \frac{2(5 \times 10^{-9} \text{ m})(0.073 \text{ N/m})(1.6 \times 10^7)}{(2 \times 10^{-5} \text{ m})^2(1000 \text{ kg/m}^3)(9.8 \text{ m/s}^2)} = 2920 \text{ m}$$

Wow. This suggests a tree *could* be 3 km tall, at least if our assumption about the number of pores is correct. Not surprisingly, other constraints intervene before this point. But this should make it clear that, despite the fact that trees have no pumps to do the job for them, the problem of moving water to the tops of trees is not a major limitation on tree size. Trees don't have to raise water. The water, taking advantage again of random thermal motion, does the work itself!

What does limit tree height? To understand this you have to imagine what's happening in the long xylem tube. Here you have a tube of water perhaps 100 meters long, suspended from a strong supporting force at the top. The column of water is supported by tension; the cohesion of the water itself. In this 'rope' of water, there isn't positive pressure squeezing the water together, but rather negative pressure, trying to pull the water apart. This tension is largest near the top, because there it must support all the water below it. Near the bottom it falls to zero.

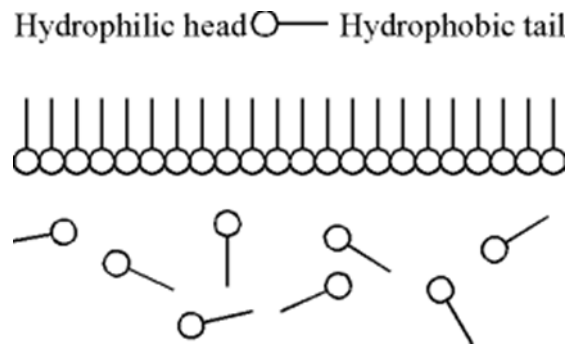
If we divide the tension in the column of water by its area we get something with units of pressure: force per unit area. This negative pressure has been measured in some of the tallest redwood trees still standing. Near the tops of these trees, it becomes as large as $2 \times 10^6 \text{ N/m}^2$, about 20 atmospheres of negative pressure.

Thin films and bubbles

Surface tension effects become important in any system where the ratio of surface area to volume is large. We have seen one example of this already: things which are small. But there is another way to make surface area large relative to volume; stretch things out into thin sheets. One place where this commonly occurs, both in our technology and in life, is in bubbles and foams. Making films, bubbles, and foams is something of a challenge. After all, we have stressed that surface tension tends to minimize surface area, and here we're talking about objects which have very much larger surfaces than they might.

In pure water, for example, making a bubble is very difficult. The surface tension is just too strong. When we want to make a bubble, we have to alter the surface tension somehow. Most often this is done by adding a small amount of another substance to the water; something called a surfactant. This word was actually invented by a company in the 1950s to refer to a whole list of chemicals which effect surface interactions of liquids, especially water. Many of these are very familiar, like soap.

Surfactants like soap alter the surface properties of water because of the unbalanced way a soap molecule interacts with water. A typical soap molecule consists of a hydrophobic tail and a hydrophilic head. When you put such molecules in water, it is energetically favorable for them to line up at the surface with their tails sticking out and their heads sticking in. They quite quickly create a layer there, excluding water molecules from the surface. Once they do that, the properties of the surface are governed by interactions among the soap molecules, rather than among the water molecules. The properties of the surface have been changed. Since it typically costs little energy to pull apart two soap molecules and let another slip in, the surface tension can be substantially reduced.

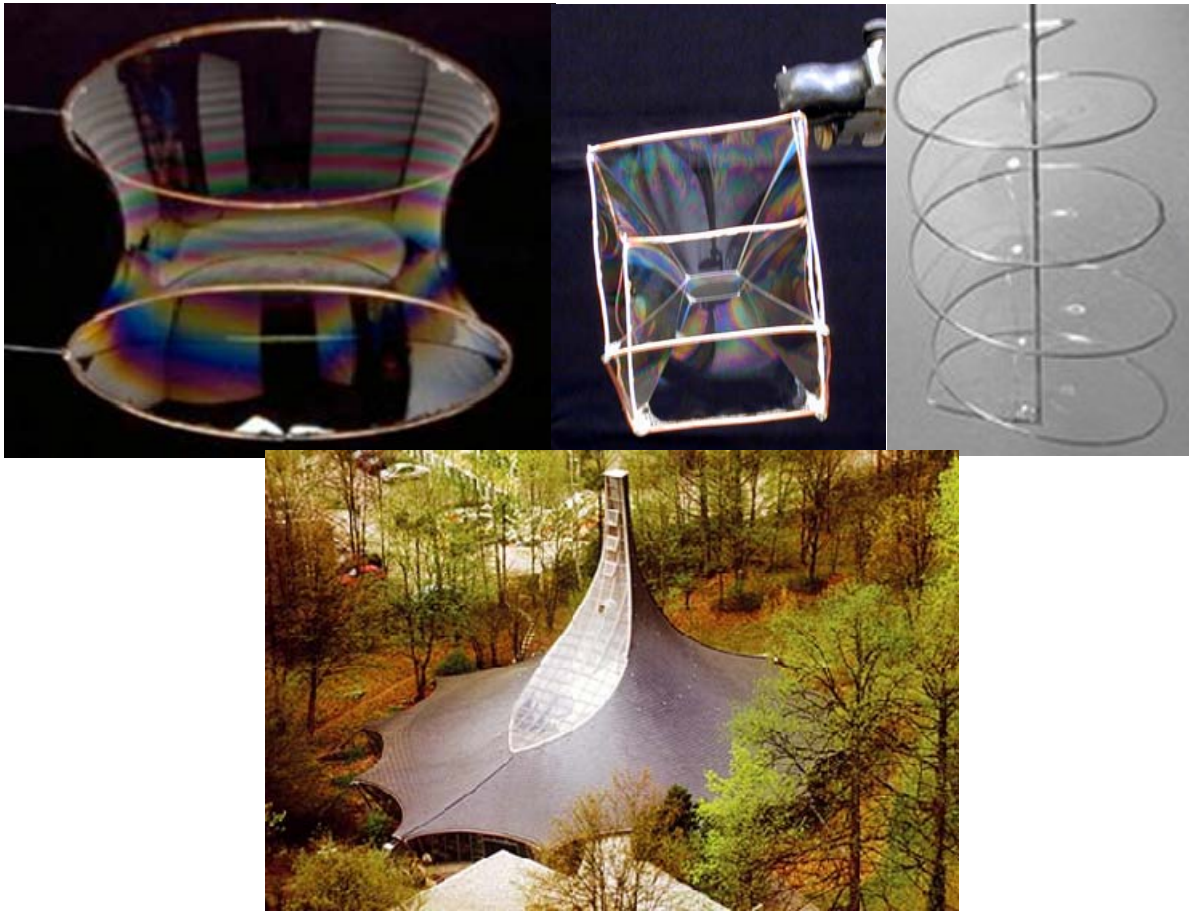


Soap films and bubbles are the most familiar example of extreme surface tension effects, and especially of the beautiful minimization phenomena which it causes. Because the presence of the soap so dramatically lowers the surface tension, it's possible to create a much larger surface and have it remain stable. This is what allows you to stretch out a small droplet of water into a large, very thin sheet.

Since life is so dependent upon water and its manipulation, it is not surprising that surfactants play many important roles in life. We will first discuss some basic properties of soap films and bubbles, relying on this familiar example to develop some useful principles. Then we will look at several applications of these principals to life.

Soap films and minimization

Because liquids with surface energy will flow until as much of that surface energy is released as possible, they will work to minimize their surface area. This minimization can be put to work for you. At the end of the 19th century a number of knotty mathematical problems about minimal surfaces were solved in part by reference to soap bubbles. These bubbles, because they sought their lowest surface energy states, often arranged themselves in truly minimal surfaces. Wouldn't the minimal surface area always be a sphere? Yes, but by attaching the fluid to solid frames like this, or by putting it into a constrained geometry, you can trap the liquid in a local energy minimum, like a bubble. With a little energy input from the outside (say from a pinprick), you can help it to jump to a still lower energy state, like the little sphere. Mathematicians seeking solutions for minimal surfaces were able to use soap bubbles as physical computers. Today, these minimal surfaces are sometimes echoed in architecture.



Bubbles: thin films trapped in a local energy minimum

Even when water is seeded with a surfactant like soap, it still has a surface tension. In fact most ordinary surfactants reduce the surface tension by a factor of three to five. Given this, we would expect any thin film of soapy water to pull itself inward, still trying to minimize its surface area, though perhaps without the enthusiasm of pure water. A way to stabilize this is to trap the film in a geometry from which it can't so easily escape: make a bubble.

A soap bubble is a thin film of water lined on both sides by layers of surfactant. Most of the film is water. When you blow a bubble, you first suspend a film of soapy water from a ring, the 'wand' that comes in the bottle of bubble fluid. This film starts flat; it has minimized its surface area subject to the constraint that it is attracted to the wand (which it wets). When you blow into the film, you add energy to it, enabling it to create new surface. When you stretch it far enough, it folds over on itself and pinches off into a bubble. Such a bubble has a very large amount of surface area given its volume. It would quite like to shrink down into a little spherical droplet. But it can't; there is an energy barrier it needs to get over because of its shape. To shrink down, this drop has to get rid of the air inside, and to do this, something needs to create a hole in the film. Punching a hole in the film requires energy; more than is present in the film itself. Your finger could do it though. Once you do punch a hole, the bubble quickly collapses, pulled inward by surface tension.



The gas pressure inside a bubble is larger than the pressure outside. It is this larger pressure which bends the soap film outward, and this is what you provide when you blow the bubble in the first place. We can find this pressure difference by considering each hemisphere of the bubble.

Thinking about one hemisphere, we can add up the forces which act on this hemisphere. First there is a force caused by the pressure imbalance, larger inside and smaller outside. The net force on one is just the pressure difference between inside and outside (ΔP) times the cross-sectional area of the hemisphere πr_{bubble}^2 . This is perhaps not completely obvious. The internal pressure pushes straight out from the center of the bubble at all points, so it has both up and down components as well as leftward ones. These sum to a total force:

$$F_{\text{pressure difference}} = \pi r_{\text{bubble}}^2 \Delta P$$

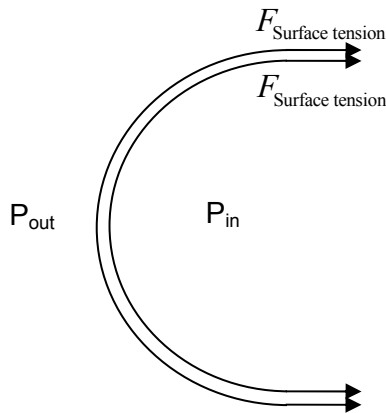
Balancing this pressure to the left is a surface tension force to the right. This surface tension force is the usual $F = \gamma L$, but we have to remember that there are two surfaces of the film, both an inner and an outer surface. The outward pressure force and inward surface tension forces balance.

$$\pi r_{\text{bubble}}^2 \Delta P = 2 \times \gamma_{\text{soapy water}} (2\pi r_{\text{bubble}})$$

or

$$\Delta P = \frac{4\gamma_{\text{soapy water}}}{r_{\text{bubble}}}$$

Notice the dependence on both the surface tension and the radius. Larger surface tension squeezes the bubble inward, so the pressure difference has to be larger to maintain the same radius. Smaller radius makes it curve more, so again the pressure difference has to be more. Imagine a really big bubble, with a very large r . The surface of such a bubble would be almost flat, and only a small pressure difference would be required to make it bend this tiny bit.



This derivation is correct for soap bubble filled with and surrounded by air. The approach also works for bubbles of gas submersed in a liquid, but there is a small difference. For a bubble submersed in a liquid, there is only one surface available to provide the surface tension. So the pressure difference for a given radius is half as large.

A big difference between surface tension and elastic tension

At first, it looks as if surface tension in a liquid is really very like the elastic tension which governs the stretching of a rubber sheet. But this is definitely not the case. Rubber sheets are elastic, which means that the force required to stretch them depends on the amount by which they are stretched. The farther you stretch them, the harder you have to pull.

This is not the case with surface tension. As there are essentially always more atoms to bring to the surface, stretching the surface a second meter always requires the same force as was required for the first meter. The force required to stretch the surface never becomes larger.

Three important applications of surface tension in life

Pulmonary surfactants: Animals living on land face an interesting interface challenge. They all transfer oxygen from their environment to their blood via diffusion. As we have seen, using diffusion effectively requires a very large surface area over which blood vessels and air are separated by only a short distance. To make such a large surface area in a relatively small volume, your lungs contain a large number of tiny, roughly spherical alveoli. There are about 600 million in the typical person, and each has a radius of around 150 microns. They have a total surface area of around 170 square meters, over 100 times the surface area of your body.

These small spheres are very like bubbles in water. To create a bubble of radius r_{bubble} requires a pressure difference of:

$$\Delta P = \frac{2\gamma}{r_{\text{bubble}}}$$

If the alveoli were lined with pure water, this would require a pressure difference of about 1000 N/m^2 , even when they're fully inflated. When you exhale and they become smaller, the pressure required to support them would normally increase substantially. Reducing the pressure required for reinflation allows the alveoli to become smaller, increasing their surface area to volume ratio. The surfactant in your lungs, generally called pulmonary surfactant, reduces the surface tension of alveolar fluid by about a factor of three at normal inflation, but their complete effect is better, and more interesting.

When you exhale and the alveoli shrink, the pulmonary surfactant occupies an even greater fraction of the surface, reducing the surface tension nearly to zero. Obviously this makes reinflation even easier; just when providing the required pressure would be most difficult, reduced surface tension makes it easy.

Surface propulsion: Quite a few organisms live right at the interface between air and water. Those which are small (mainly arthropods, including many insects) take advantage of surface tension in many ways. For large objects, the forces associated with surface tension are easy to ignore. For insects, they can be quite substantial. One way to cheaply move around a water surface is to create a surface tension gradient. Imagine that the surface tension in front of a little beetle is larger than the surface tension behind. This unbalanced force would pull it along the surface with no effort on its part.

Not too surprisingly, a number of surface living insects use this approach. Most of them do this by excreting a surfactant into the water behind them. This lowers the surface tension behind them, while not altering it in front of them, and allows the surface tension of the water to zip them across the surface. When the insect is small, this can be very effective, often moving them many times faster across the surface than they could walk. You can emulate this method with a piece of a toothpick. Just put a little soap on one end and drop it into some water still water. Remember, any soap which hits the water will spread out over the surface very quickly, and once it does, the effect will disappear.

Meniscus climbing: If you live on the surface of water and would like to climb out onto land, the meniscus at the edge of the water can be a serious barrier. Quite a number of insects have hydrophobic feet so that they can walk on top of the water. Many of these deal with meniscus climbing by deploying wettable extensions to their forelegs. Once the forelegs have adhered to the surface, surface tension can pull little water walker right up the meniscus. These and many other examples of creature making their



way on the surface of water are described in a recent review by MIT mathematicians Bush and Hu.¹

Relation to solids and practical importance

There are many surface effects in solids which are analogous to the surface effects in liquids. The real difference of course, is that new atoms don't move to the surface on their own: simply due to thermal motion. But very similar statements can be made about why solids don't fall apart. The bonded material is energetically favored over more surface area, so to break it you have to apply energy. In order to split a diamond, you have to provide enough energy to account for the new surface energy of the two pieces you create.

Surface physics, the study of the details of what goes on right at the surface (primarily of solids) is a major field of study today. Why would the study of surfaces become more and more important as time goes on? In this age of minimization everything that we work with becomes more "surface-dominated" all the time. So like soap bubbles, the tiny structures used in the modern electronic industry are more and more dominated in their behavior by surface effects. The ratio of surface area to volume changes like $1/r$, so as the size of things gets small, the relative importance of surface phenomena increases.

When does a material become "mostly surface"? These are things called mesoscopic, they are between macroscopic and microscopic, and they have interesting properties. Once the radius of the atoms becomes a modest fraction of the total size of the object, the fraction of atoms which are at the surface becomes large, and surface effects become important.

Much effort now is going into "nanotechnology", the manufacture of very tiny, subcellular technologies. Nanotech devices are very much in this surface dominated regime. That's part of the reason why they offer new opportunities for engineering.

A Quick Summary of Some Important Relations

Hydrostatics:

To support the weight of the fluid above, fluid pressure must increase with depth.

$$P_d = P_{\text{top}} + \rho g d$$

Pascal's principle says that an increase in the external pressure on a fluid is transmitted equally throughout; this is used in hydraulic force systems as a kind of force multiplier.

Buoyancy:

For any object even partly submerged in a fluid there is a buoyant force

$$F_{\text{buoyant}} = \rho_{\text{fluid}} V_{\text{object in fluid}} g = \text{weight of fluid displaced}$$

For a floating object, buoyant force balances weight:

$$\frac{V_{\text{in floating}}}{V_{\text{total}}} = \frac{\rho_{\text{object}}}{\rho_{\text{fluid}}}$$

Cohesion, surface energy and surface tension:

There is extra energy associated with atoms at the surface of a liquid. This energy, typically small, is important when the SA/Vol ratio is large – for small things. Surface energy generates a surface tension:

$$F_{\text{surface tension}} = \lambda L$$

Surface energy and wetting:

The ability of a liquid to spread on a surface, to 'wet' it, depends on a balance between cohesion in the liquid and adhesion between the liquid and solid. This balance can be expressed by a contact angle. When this angle is less than 90°, the liquid tends to spread and wet the surface. When it is greater than 90°, the liquid tends to 'ball up' on the surface, and not wet it.

Thin films, surfactants, and bubbles:

Surfactants are chemicals which can dramatically alter surface properties even at small concentrations. When they are used to lower the surface energy of water they allow the formation of thin films like soap bubbles. Soap bubbles and other thin film bubbles have a pressure difference from inside to outside:

$$\Delta P_{\text{thin film bubble}} = \frac{4\gamma}{r}$$

Bubbles of air in water have only one surface, and hence have a smaller pressure difference between inside and out:

$$\Delta P_{\text{gas in fluid bubble}} = \frac{2\gamma}{r}$$

Surfactants play an essential role in controlling the pressure required to expand and contract the alveoli in your lungs.

ⁱ Bush, J., and Hu, D., 2005, “Walking on Water: Biocomotion at the Interface”, Annual Reviews of Fluid Mechanics. **38**, 339.

18. Flowing: fluid dynamics, including viscosity and flow in a pipe

- 1) Fluid flow and idealization
 - i. Continuity
 - ii. Continuity analysis of the circulatory system
- 2) Energy in flows
 - i. Three forms to start: Kinetic, gravitational potential, and elastic potential
 - ii. The Bernoulli equation and flows
 - iii. Accounting for losses: thermal energy
- 3) Approaching real flows: Newtonian viscosity
 - i. Viscosity and the time rate of change of strain
 - ii. Key consequences of viscosity
 - iii. Limitations of this picture: non-Newtonian fluids
 - iv. Kinematic viscosity and flow stability
- 4) Characterizing flows on different scales
 - i. Reynolds' number: how to estimate it
 - ii. The meaning of Reynolds' number
 - iii. Implications of scale dependence
- 5) Life at large and small Reynolds' number
 - i. Fluid mixing is typical and motions are inertia dominated
 - ii. Fluid mixing is absent and motions through fluids are friction dominated
- 6) Reynolds' number and terminal velocity
 - i. The Stokes regime
 - ii. The Rayleigh regime
 - iii. Intermediate cases

Physics for the Life Sciences: Chapter 18

The key property of fluids, what makes them so interesting, is that they flow. The atoms which make up the fluid are able to move relative to one another, rearranging its structure under even the most subtle of influences. This ability to change, to mix and reshape, is why fluids are so important for life. In a solid, nothing much ever happens. All of its atoms remain fixed in place, oscillating a bit, but never rearranging themselves. Life requires continuous change as it feeds off of the flow of energy through it. This could never happen without the freedom of change which fluids provide.

All living things are immersed in fluids, either air or water, and move relative to them. This may happen because of locomotion (a bird flying or fish swimming), because the fluid flows by (a tree in a storm), or because the fluid flows through the organism (a circulatory or respiratory system). Understanding how fluids flow is essential for understanding life. We will approach this subject in steps. First, we'll consider the steady flows of "ideal" fluids. As usual, this simplified model will allow us to extract some important general principles. We will see how to relate the flow rate in one location to that in another, and understand how energy contained in a flow can be passed from one form to another as it moves.

After exploring these basics, we will introduce the most important realistic features of fluid flow. Friction will be the first. It has two principal effects, converting the macroscopic energy of the flow to thermal energy, and structuring the profile of a flow, preventing large velocity gradients from persisting. The

second complication is turbulence and mixing, which take the ability of a fluid to rearrange its constituents to an extreme.

18.1 Fluid Flow, and the necessity for idealization

What kind of fluid would we call ideal? An ideal fluid flows in a way which is completely free and uncomplicated. Such a fluid should always flow to escape pressure, rather than be compressed and push back for example. It should present no other resistance to flow either; it should flow without friction and not swirl around and get mixed up when it does. These two constraints are much the same.

To develop our first conclusions about fluid flow we will imagine that the following limits apply:

- The flow is steady. This doesn't mean that it is the same everywhere, but only that at each point the flow is not changing with time. In a sense this means we will be studying the statics of fluid flow. This condition implies that the shear stress on the fluid should be constant.
- The flow is laminar. A laminar flow is layered, so that it doesn't mix. Each bit of fluid follows the bit in front of it along a 'streamline' and these streamlines never cross.
- The fluid is incompressible: its density doesn't change. This is a good approximation for water, but quite poor for air.
- The fluid should have no internal friction. This means no energy is lost in the continuous flow of the fluid. Later we will express the friction which a real fluid experiences as viscosity, and describe a fluid without friction with the delightful adjective 'inviscid'.

In modeling complex phenomena, we often make assumptions like this; we ignore air friction, treat things as point particles, pretend that objects are rigid and don't deform. Each time we do this we make errors, and we need to be careful to understand what they might be.

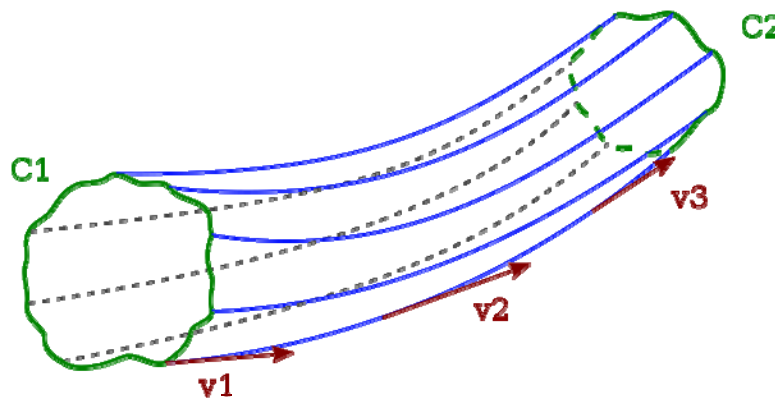
We reiterate this topic here because the assumptions of ideal fluids and their flow are *very* rarely met. These simplifying assumptions, in a practical sense, are a much worse than many we have made. Nevertheless, assuming these things allows us to understand two simple principles which, while they may not be obeyed quantitatively, provide very good qualitative descriptions of what happens in fluids.

Simplest property of fluid flow: continuity

The first principle of fluid flow we will use is called the continuity equation. It concerns the flow of the fluid through a confined channel, something like a pipe, and makes a simple assertion; what goes in must come out. The volume flow rate is a measure of the volume of fluid V which flows through a surface per unit time. This is related in a simple way to the cross-sectional area perpendicular to the flow (A) and the speed of the flow (v):

$$\frac{dV}{dt} = A \frac{dx}{dt} = Av$$

If I make the channel wider while keeping the flow velocity constant, the flow rate increases. If I increase the velocity while keeping the area constant the flow rate increases. This volume rate of flow is also sometimes called the 'flux' of the fluid.



If the fluid is incompressible, then any fluid which passes through a first surface in the flow A_1 must also pass through a later surface A_2 . The continuity equation expresses this. It just says that this flow rate must be the same at different locations in a confined flow:

$$A_1 v_1 = A_2 v_2$$

This is just saying that total amount of fluid passing each point along the flow is the same. If this weren't true the fluid would have to pile up somewhere along the way.

Being a little more careful, the "area" used here should be perpendicular to the streamlines of fluid flow. This principle depends on the incompressibility of the fluid. If it is compressible, then it can change its volume, and so the volume rate of flow need not be conserved. For liquids at least, incompressibility is not such a bad approximation. For gases it is a very bad one indeed.

A simple application of the continuity equation

Here is a simple, and interesting, application of the continuity equation. Your circulatory system is a remarkable network, using fluid flow to deliver oxygen and nutrients to all the parts of your body. To accomplish this, it must send all the blood through a single channel in your heart, then split the flow out into many branches, sending blood off into ever narrowing channels, until finally at least one tiny capillary passes by every living cell in your body. It would be interesting to know just how many capillaries you have. But this is *really* hard to determine by counting.

Let's use the continuity equation to get some idea of the number of capillaries you have. The aorta, the major blood vessel coming out of the heart, has a radius of about 1.5 cm, and blood flows through it at an average rate of about 30 cm/s. It's not especially hard to observe the blood flow in a single capillary, and measure its rate. If blood flows through the capillaries at about 0.04 cm/s, what is the total area of capillaries in your body?

We can answer this using the continuity equation.

$$A_{\text{aorta}} v_{\text{aorta}} = A_{\text{capillaries}} v_{\text{capillaries}}$$

$$A_{\text{capillaries}} = A_{\text{aorta}} \left(\frac{v_{\text{aorta}}}{v_{\text{capillaries}}} \right)$$

Putting in the numerical values for these things we find:

$$A_{\text{capillaries}} = (\pi r_{\text{aorta}}^2) \left(\frac{v_{\text{aorta}}}{v_{\text{capillaries}}} \right) = (7.1 \times 10^{-4} \text{ m}^2) \left(\frac{30 \text{ cm/s}}{0.04 \text{ cm/s}} \right) = 0.53 \text{ m}^2$$

Now that's a quite a big cross-sectional area. A circular pipe with that large an area would have to have a radius of about 16 inches! When your heart pumps blood out of your heart, it is sending it through a channel which starts smaller than a garden hose then broadens out into something with the area of a sizable sewer pipe, then narrows down dramatically again to reenter your heart. Along the way, the flow slows by about a factor of 750. Blood flows through the capillaries really slowly. Normally to get an idea of a velocity, we might convert it to miles per hour. This capillary flow rate is 0.0009 miles per hour, which isn't too helpful. So it might be better to convert this and recognize that something going this fast would cover one mile in just about three years. Truly slow, and the kind of speed you typically encounter only on microscopic scales. I leave it to you to figure out how blood traveling so slowly can ever make it back to the heart.

Since capillaries have typical radii of 5×10^{-6} m, the total number of capillaries separating your arteries and veins is about:

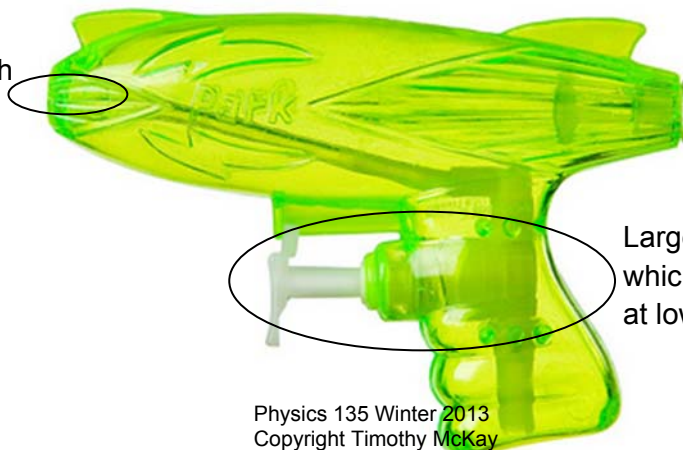
$$N_{\text{capillaries}} = \frac{A_{\text{capillaries total}}}{A_{\text{one capillary}}} = \frac{0.53 \text{ m}^2}{8 \times 10^{-11} \text{ m}^2} = 6.6 \times 10^9$$

The flow which starts in one large aorta splits out into something like six billion capillaries. These capillaries service something like 5×10^{14} (tens of trillions) cells in your body. So the very large number of capillaries is not so surprising.

This same principle is at work in the summer play associated with water. A squirt gun converts a relatively low velocity flow in a large piston to a high velocity flow through a narrow opening.

This is also what happens when you cover a part of the end of a hose with your thumb. To pass through the much smaller area you leave open, the water must travel much faster.

Small hole through which fluid flows at high velocity



Large piston in which fluid flows at low velocity

18.2 A somewhat more complex fluid flow relation: energy conservation

In an ideal fluid, there is no loss of energy to friction, and the bulk mechanical energy in a flow of fluid is conserved. That is, there's no conversion from bulk mechanical energy into random thermal energy. When a fluid flows along, it contains some amount of energy per unit mass. That energy might be in any of the several forms which an extended object can have; typically kinetic energy, gravitational potential energy, and elastic potential energy. Just as when we throw a ball up it exchanges kinetic energy for potential, so too a fluid trades energy among different forms as it flows.

When water flows in a pipe, each bit of water possesses some energy. What are the forms this energy can take?

- Kinetic energy (bulk fluid flow)
- Gravitational potential energy (height of fluid)
- Elastic energy (how much is it squeezed, pressure)
- There would be thermal energy too, only we specified no friction, which means that energy does not flow from bulk mechanical forms to random thermal forms for such an ideal fluid...

To discuss these within the flow of the fluid we need to know how much of each kind of energy the fluid possess *per unit volume*. What is the specific energy in each bit of the flow? For kinetic energy this is simple, we just imagine a small subset of the flow with mass Δm and volume ΔV and we have:

$$\frac{\text{Kinetic energy}}{\text{Unit volume}} = \frac{\frac{1}{2} \Delta m v^2}{\Delta V} = \frac{1}{2} \rho v^2$$

Gravitational potential energy is similar. If we agree to measure this relative to some reference point a distance h below the location of this bit of fluid, it has gravitational potential energy per unit volume given by:

$$\frac{\text{Gravitational potential energy}}{\text{Unit volume}} = \frac{\Delta m g h}{\Delta V} = \rho g h$$

The elastic potential energy per unit volume is perhaps a little less obvious. As it turns out, this quantity is measured directly by the pressure P . We can see this by imagining the energy the fluid would give up if we allowed it to expand. If this little volume of fluid expanded by an amount dV , it would do an amount of work on its surroundings given by

$$dW = P dV$$

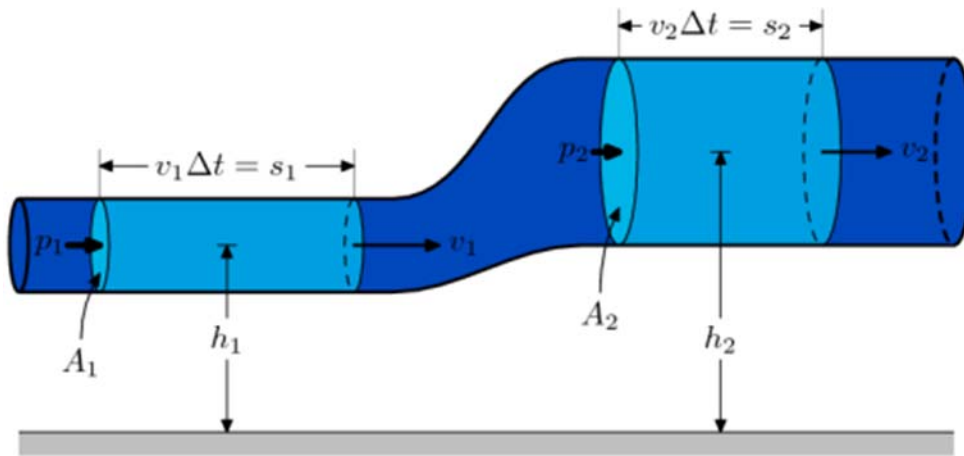
This work would represent potential energy taken from the fluid, and from this we can see that pressure measures the amount of energy available in elastic form:

$$P = \frac{dW}{dV}$$

We expect to see the total amount of energy in these three forms conserved as the fluid flows from one place to another. A little bit of fluid contains the same total amount of energy as it flows along, though it may be kinetic energy in on part of the flow and potential energy in another. Comparing the mix of these three forms at two different locations in the flow we could write:

$$P_1 + \frac{1}{2} \rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2} \rho v_2^2 + \rho g h_2$$

This description of energy conservation in a fluid flow is called Bernoulli's principle, and is named for the mathematician Daniel Bernoulli, who developed a form of it in 1738. For a simple proof consider a tube of fluid flowing as shown in the picture below.



Imagine what happens to the bit of fluid in the lower part of the pipe (in the region labeled 1). This fluid has a total mass

$$m = \rho A_1 v_1 \Delta t$$

After this piece of the fluid moves to the upper part of the pipe, it will have some new velocity v_2 . We can determine the change in kinetic energy of the fluid from the work-energy theorem

$$W_{\text{total}} = W_{\text{gravity}} + W_{\text{pressure}} = \frac{1}{2} m (v_2^2 - v_1^2)$$

We can calculate each of these work contributions

$$W_{\text{gravity}} = -mg(h_2 - h_1)$$

$$W_{\text{pressure}} = -(P_2 A_2 s_2 - P_1 A_1 s_1) = -(P_2 - P_1)V$$

So that:

$$-(P_2 - P_1)V - mg(h_2 - h_1) = \frac{1}{2} m v_2^2 - \frac{1}{2} m v_1^2$$

which can be rearranged as:

$$P_1 + \frac{1}{2} \rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2} \rho v_2^2 + \rho g h_2$$

This is just a statement of conservation of energy for an incompressible, inviscid fluid-flow.

What the Bernoulli principle says is that for *ideal* fluids, energy is conserved. Any little piece of the fluid, as it moves along, will convert its energy from one form to another, but will neither gain nor lose energy. So, I can assume that the total energy per unit volume will be the same throughout the flow. Measure it at one place, and it will be the same someplace else. But it is important that you just remember what this really means, rather than just using the equation. It tells us to find the energy per unit volume somewhere in the flow, and then we know it will be the same somewhere else.

Examples of applying the Bernoulli equation

Water pressure in your house: How does the water pressure change in your house? Think about how fast fluid comes out of the pipe on the first and second floors. In both cases the fluid coming out of the pipe is at atmospheric pressure. It must be because there's nothing else to prevent it from flowing. So:

$$P_{\text{atm}} + \frac{1}{2} \rho v_1^2 + \rho g h_1 = P_{\text{atm}} + \frac{1}{2} \rho v_2^2 + \rho g h_2$$

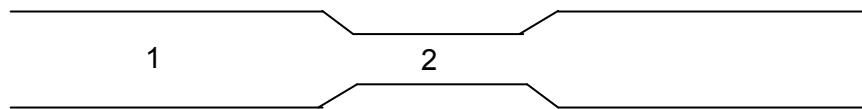
Or

$$v_2^2 = v_1^2 - 2g(h_2 - h_1)$$

The bigger the change in height, the lower the velocity with which it emerges; too bad if you live on the fifth floor of a dorm. The water may well dribble out of your shower, even though it comes screaming out down on the first floor.

Flow in a restricted pipe:

Imagine the flow of water in a pipe which is wide in one place and narrow in another. This is a model for the squirt gun described above. It might also represent what happens when arteriosclerosis.



Here we again begin with Bernoulli equation, comparing the flow at point one and point two. Now the height of the two points is the same, so we have:

$$P_1 + \frac{1}{2} \rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2} \rho v_2^2 + \rho g h_1$$

So what we find is that the pressure changes as the flow goes from point one to point two:

$$P_1 - P_2 = \frac{1}{2} \rho (v_2^2 - v_1^2)$$

Remember, we could tell how the velocities changed from the equation of continuity:

$$A_1 v_1 = A_2 v_2 \quad \text{or} \quad v_2 = \left(\frac{A_1}{A_2} \right) v_1$$

This can let us rewrite the above as:

$$P_1 - P_2 = \frac{1}{2} \rho \left(\left(\frac{A_1}{A_2} \right)^2 - 1 \right) v_1^2$$

Now since $A_1 > A_2$, we know that $v_2 > v_1$. And from the equation we just wrote, we see that $P_1 > P_2$.

When the fluid flows faster, the pressure is reduced. This should be an obvious consequence of the Bernoulli principle. If the fluid is going to speed up, the energy which is required has to come from somewhere. It can't come from gravitational energy, as the flow is horizontal. So it must come from the elastic energy stored as pressure.

Water leaking from a tank: If a tank, open to the air on top, has a leak a distance d_{hole} below the surface, how fast does the water come out?

At the top of the fluid the pressure is atmospheric, the velocity of the fluid is essentially zero, and the height is a distance d_{hole} above the hole. At the hole, we might take the height to be zero, the velocity is unknown, and the pressure is again atmospheric pressure. Why is this? When a bit of fluid is right at the hole, the only pressure holding it in is atmospheric pressure. So we have:

$$P_{\text{atm}} + \rho g h_{\text{hole}} + \frac{1}{2} \rho v_{\text{top}}^2 = P_{\text{atm}} + \rho g (0) + \frac{1}{2} \rho v_{\text{hole}}^2$$

Or

$$v_{\text{hole}} = \sqrt{2gd_{\text{hole}}}$$

We could examine this case a little further by considering three points in the tank: one at the top, one at the depth of the hole, but still inside the tank, and the third at the hole. Following a little fluid from one to the next, we find that it starts with some gravitational potential energy. Then as it moves deeper, still going very slowly, most of the original energy is converted to pressure. Then as it approaches the hole, the pressure it has gained when going deeper is converted to kinetic energy as it squirts out.

Limitations

In these examples we have seen some of the ways in which energy can be converted from one form to another within a fluid. In the first we saw how kinetic energy could be used up creating potential energy as water flows higher in your house. In the second we saw how pressure could be converted into kinetic energy in a restricted flow. In the third we saw how gravitational potential energy could be turned into first pressure, and then kinetic energy, as water flows from a leak in a tank.

How good is the Bernoulli equation? What would happen if, during the flow, some of the energy was lost due to friction? In practice, this would change the Bernoulli equation into an inequality. The flow would still begin with some combination of pressure, gravitational potential energy, and kinetic energy. At a later point in the flow it might still have all of these, but total amount of energy present in these forms would be reduced. So if point one is earlier in the flow than point two, we might rewrite the Bernoulli equation as an inequality like this:

$$P_1 + \rho gh_1 + \frac{1}{2} \rho v_1^2 > P_2 + \rho gh_2 + \frac{1}{2} \rho v_2^2$$

Or as an equality like this:

$$P_1 + \rho gh_1 + \frac{1}{2} \rho v_1^2 = P_2 + \rho gh_2 + \frac{1}{2} \rho v_2^2 + (\text{Energy lost to friction})$$

You will not be surprised to learn that energy present in the form of random thermal motion is difficult to convert back into motion in a fluid flow.

Let's consider a simple example in which we allow for this realism. Imagine that you have a long, horizontal, uniform pipe through which you want to push some water. You put it in one end with pressure P_1 and velocity v_1 . It comes out the other end with pressure P_2 and velocity v_2 . If this fluid were ideal, everything would remain exactly the same, and this flow would occur with no energy loss. If the fluid is not ideal, then we would find from the equation above that:

$$P_1 + \frac{1}{2} \rho v_1^2 = P_2 + \frac{1}{2} \rho v_2^2 + (\text{Energy lost to friction})$$

Continuity requires that $v_1 = v_2$, so we can rewrite this as:

$$P_1 - P_2 = (\text{Energy lost to friction})$$

To push a real fluid through a pipe you make the pressure higher at one end than the other. The energy present at the start of the flow as higher pressure is converting during the flow into thermal energy in the fluid. It costs something to make the fluid flow. We will model this energy loss later in this chapter, and write down an equation for estimating this pressure loss quantitatively. But you know it must be there just from simple considerations like this.

Consequences of Bernoulli's principle

The basic notions of Bernoulli's principle are seen in a wide range of circumstances relevant for life. The pressure differences generated by moving fluids are used by many organisms to induce flow. In many cases, organisms use the fact that fluid velocities increase with distance from a solid surface. For example, wind speed is low very close to the ground is small compared to that higher in the air. You may have experienced this at the beach. While lying on your towel, you get quite hot in the sun. But when you stand up you discover a nice cool breeze is blowing.

Sponges are an incredibly common and ancient form of sea life. They are very simple animals, mostly staying still where they grow. Rather than have a complex set of internal organs, sponges simply maintain

a constant flow of water through their pore filled bodies. They capture food and extract oxygen from this flow, dumping their waste along the way. To live comfortably, a sponge just needs to keep the fluid flowing. They make this goal simpler by extending their tops well above the sea floor, into the more rapidly flowing fluid generally found there. This makes the pressure at the top of the sponge lower than at its base, driving a flow of water in through the sides of the sponge and up through its center. As the water passes through it, tiny cilia strain it for food.

On land, prairie dogs in the American West use a similar strategy. Their very large networks of underground burrows need ventilation. The CO_2 which they exhale is denser than the O_2 they need to breathe in. If they didn't replace the O_2 , they would suffocate. The burrows are much too large for diffusion to replace the oxygen, so they need to use forced convection. To accomplish this they build a mound at one end of the system which sticks up from the ground. The mounds at the other end are lower and more rounded. This allows the pressure difference between the two ends to drive a flow of fresh air from the low mound through the high. Termite mounds, built up from the ground, perform a similar function.



18.3 Approaching real flows: Newtonian viscosity

As we have mentioned several times, real fluids are more complex than the ideal fluids described by the continuity equation and Bernoulli's equation. John von Neumann, one of the great early physicist/computer scientists, once commented that assuming no friction in fluids is like "working with dry water". Fluid dynamics is one of the most complex subjects in physics, and is now often addressed using the world's largest supercomputers. Many of the important and interesting phenomena associated with fluids emerge from the "extra" properties of real fluids. As you might expect, these interesting properties of real fluids are also regularly used by living things to accomplish their goals, so it is especially important that you should understand some aspects of them.

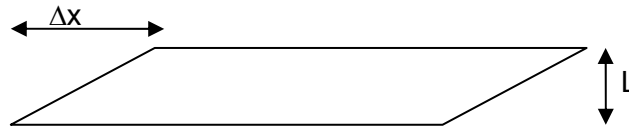
In this section we will introduce real fluids, and at least get an idea of how we describe them. Along the way we'll see how even a basic consideration of friction in fluids explains a lot of new phenomena. This is a rich and very current field, important because most of the matter in the universe is in fluid form. Even the Earth itself, the canonical solid object, is mostly a fluid, covered by a thin solid crust. In addition, all life we know moves in fluids, air or water, so the behavior of these real fluids places strong constraints on how living things evolved.

Returning to the definition of fluids and shear stress

Remember how shear stress works in solids:

$$\frac{F}{A} = S \left(\frac{\Delta x}{L} \right)$$

In solids, the shear stress, the force per unit area creating the shear, is proportional to the shear strain, how much the material distorts under the stress. See the picture below:



This is what happens when you apply a shear stress in a solid. What happens in a fluid? If I apply a shear stress, instead of just distorting and coming to a halt like the solid does, it begins to flow. There is no limit to the shear strain which can be produced in a fluid. *Any* shear stress can produce an *infinite* shear strain.

So, what's the appropriate thing to associate with the varying properties of fluids? If I apply a shear stress to some water, and apply the same stress to tar, they both will flow, and both will flow as far as I like if I wait long enough. It's clear that something is different about these two fluids. How should we quantify this difference?

Viscosity and time rate of change of fluid shear.

Fluids with low internal friction shear rapidly, those with high internal friction shear slowly. This can be expressed in the equation:

$$\frac{F}{A} = \eta \left(\frac{dv}{dy} \right)$$

here we see again the shear stress. But now instead of a strain ($\Delta x / L$), we have a time rate of change of a strain (v / L), and instead of the shear modulus S we have the "viscosity" η . Viscosity is a measure of the internal friction of fluids, how much they resist flow. The units of viscosity can be found from its definition:

$$\eta = \frac{\left(\frac{F}{A}\right)}{\left(\frac{v}{L}\right)} \text{ or } \frac{\left(\frac{N}{m^2}\right)}{\left(\frac{m/s}{m}\right)} \text{ or } \left(\frac{N}{m^2}\right)s \text{ or Pascal} \times \text{second}$$

Unfortunately this SI unit is not always used in practice, and instead viscosity is often tabulated in terms of “poise”, a unit named for the French scientist Poiseuille, and defined as:

$$1 \text{ poise} = 0.1 \frac{\text{Ns}}{\text{m}^2}$$

Worse still, viscosities are often printed in tables in units of “centipoise”, or 100ths of a poise. So you’ll have to be careful if you take viscosity values from the literature. Here are some example values for viscosity measured in the SI unit of Ns/m².

Temp °C	Water	Air	Mercury	SAE 10 motor oil	SAE 30 motor oil	Glycerin	Honey
0	1.8x10 ⁻³	1.7x10 ⁻⁵	1.7x10 ⁻³				
20	1.0x10 ⁻³	1.8x10 ⁻⁵	1.6x10 ⁻³	7x10 ⁻²	0.3	0.5	1.5
40	0.7x10 ⁻³	1.9x10 ⁻⁵	1.4x10 ⁻³				
60	0.5x10 ⁻³	2.0x10 ⁻⁵					
80	0.4x10 ⁻³	2.1x10 ⁻⁵					
100	0.3x10 ⁻³	2.2x10 ⁻⁵	1.2x10 ⁻³				

Note a few facts which are apparent from this table. Liquid viscosities decrease with temperature, while gas viscosities increase. This is because the increasing thermal energy available as temperature rises in liquids makes it easier for atoms to flow past one another. In a gas, like air, the increasing temperature implies more frequent collisions among the atoms, and hence greater viscosity. It’s also worth noting that fluid viscosities, even in this short table, can vary a lot. Honey is 100,000 times as viscous as air.

Fluids which obey this linear relation between shear stress and shear velocity are called ‘Newtonian’ fluids.

Consequences of viscosity

How do fluids with viscosity flow differently from ideal fluids, which would have a viscosity of zero? Here are a few of the key consequences.

Energy is lost as a fluid flows: This effectively adds another term to Bernoulli's equation as has been discussed above. It leads to pressure losses along horizontal flows.

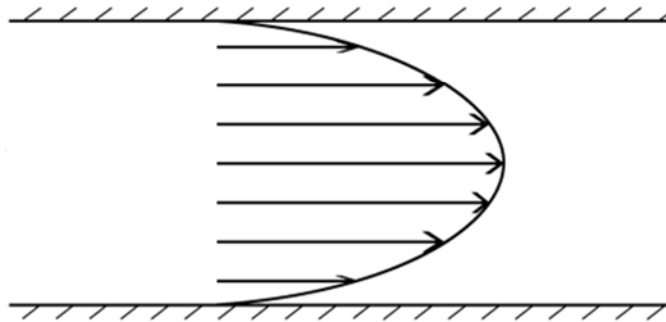
Flow layering and laminar flow: The definition of viscosity suggests that the velocity of a fluid flow will increase as you move farther into the flow. It's useful to consider the flow in a cylindrical pipe. Right at the wall of a pipe the fluid velocity is zero. This somewhat surprising fact is called the 'no-slip condition', and it is an excellent approximation in a wide variety of circumstances. Moving away from the pipe, the flow rate rises to the center of the flow, and then falls back to zero at the opposite wall.

Imagine water flowing down a river. The shear stress here comes from gravity is the same everywhere, so from the definition of viscosity you can show that:

$$v = \left(\frac{F}{\eta A} \right) \int dy = \left(\frac{F}{\eta A} \right) L$$

where L is the distance from the river's edge or bottom. The speed of the flow is proportional to the distance from the wall. This should be a familiar reality; the flow midstream is greater than at the edge. This layered, "laminar" flow is a widespread phenomenon for smooth flows, and plays an important role in many living systems.

Flow in a pipe: The no-slip condition and presence of viscosity imply that the flow velocity in a pipe will vary from small at the edge to faster in the center. The flow profile is roughly parabolic, increasing rapidly near the edge and more slowly in the center. A diagram of this velocity profile is shown below.



For laminar flows, several important relations can be derived to describe the flow rate. As we have seen, flow in a viscous fluid is driven by a pressure difference, from high pressure at the input of the pipe to lower pressure at the output. The speed of the flow at any point, as well as the total volume rate of flow, will depend on this pressure difference.

The first relation gives us an estimate of the maximum flow velocity at the center of a pipe:

$$v_{\text{center}} = \frac{r^2}{4\eta} \frac{\Delta P}{L}$$

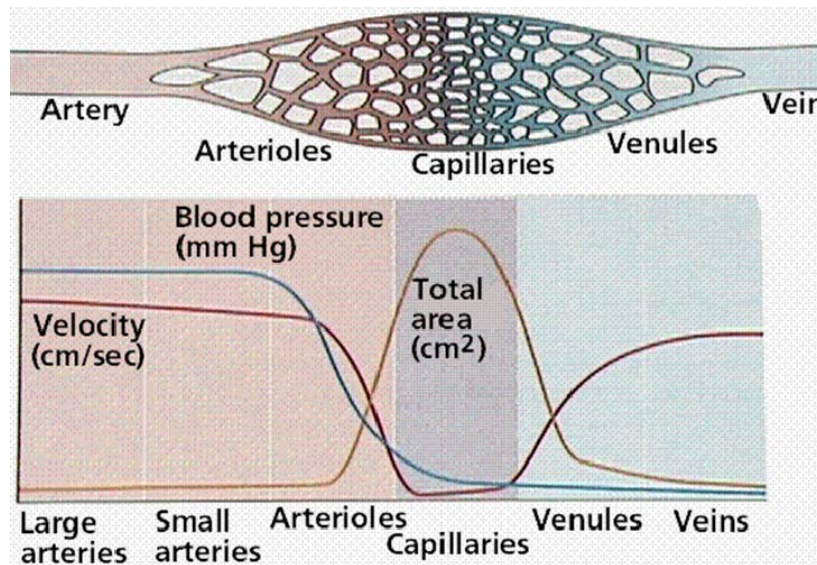
In this equation, we see that the maximum velocity depends on the radius of the pipe squared, the viscosity, and the pressure gradient $\Delta P / L$. Note that this is a gradient driven process, rather like diffusion. The second relation, called the Hagen-Poiseuille equation, describes the volume rate of flow of the fluid:

$$\Phi_{\text{whole pipe}} = \frac{dV}{dt} = \frac{\pi r^4}{8\eta} \frac{\Delta P}{L}$$

When such a viscous, incompressible fluid flows through a cylindrical pipe, the total volume rate of flow depends on the radius of the pipe to the fourth power, the viscosity, and the pressure gradient. The very strong dependence of flow rate on radius has important consequences. When an artery becomes partially blocked maintaining the flow requires larger and larger pressure gradients. If the artery's radius is reduced by a factor of two, the pressure gradient must increase by a factor of eight to maintain the same flow.

Pressure in the human circulatory system: What we have said above about flows in pipes all plays out in the human circulatory system. Blood leaving the heart is at high pressure, flowing rapidly through a relatively large artery. Pressure losses are small in this flow, because of the large radius. When the blood enters capillaries, the flow speed falls dramatically, and large pressure gradients are required to keep the flow moving. As it reemerges from the capillaries and enters large veins, its flow is again easy, and only small pressure gradients are required to return the blood to the heart.

The changes of pressure, local velocity, and total cross-sectional area within the circulatory system are sketched in the figure below.



Resistance to the motion of small objects through a fluid: The rate at which the fluid can move out of the way of a small object moving slowly is dependent on its viscosity. One example is a falling object. Under a constant strain (the weight of the falling body) the fluid reaches a constant velocity relative to it, the value of which is governed by the viscosity η .

This is actually a bit more complicated. Even motion through a fluid without friction will exhibit terminal velocity. Just accelerating the fluid to get it out of the way requires that a falling object exert a force on the fluid, implying a reciprocal force on the falling object. For ‘large’ objects falling at ‘large’ velocities, like you or I falling through air, this inertial effect is the most important. For small objects falling at low velocities, the frictional effects associated with viscosity are more important.

You can see that in the two fluid friction laws we discussed much earlier in the class. The small-slow form of friction depends on the viscosity of the fluid, but not its density, while the large-fast form depends on the density of the fluid, but not its viscosity:

$$F_{\text{fluid friction}}^{\text{small-slow}} = 6\pi\eta rv$$
$$F_{\text{fluid friction}}^{\text{large-fast}} = \frac{1}{2}C\rho Av^2$$

We will (finally!) see in a more quantitative way how to decide what constitutes large-fast and small-slow objects below.

Limitations of this picture

This picture of Newtonian fluid behavior is a kind of first approximation, much better than assuming no friction. But it is limited. The Newtonian fluids we are describing here have viscosity η which is independent of velocity.

There are many non-Newtonian fluids, including most biological fluids. Blood provides a good example. As the velocity of blood increases, cells align themselves with the flow, lowering the friction. Such a material is called “shear-thinning”, because it becomes less viscous once it starts to flow. This is common in suspensions and otherwise non-uniform liquids.

Many other common examples of non-Newtonian fluids exist, like egg white (albumen), a “viscoelastic” fluid, which is elastic for small shears and can spring back to its original shape, but will flow like any liquid for larger shears. Ketchup is another well known shear-thinning fluid. Once ketchup starts to flow, it does so very freely, blopping out all over your French fries. Then there is the famous corn starch and water concoction you may have played with in elementary school. This is a “shear-thickening” fluid. If you apply a small shear it will flow nicely, but if you apply a large shear it becomes very viscous indeed.

The analysis we have done here works best with uniform, one component fluids. This implies an opportunity for life. If it wants to get around the limitations of simple fluids, it can often work up a kind of a cocktail mix of things which will behave the way it needs. Obe recently published example describes the viscoelastic properties of the digestive juices of a carnivorous pitcher plant. This gooeyness greatly increases the chance that a fly landing in the fluid will be unable to escape¹

Kinematic viscosity

So far all of our discussion has concerned fluid flows which are static. We’ve been discussing the **stable** state of flow which the fluid reaches after we give it time to settle down. The dynamical response of a

¹ Gaume L, Forterre Y 2007 A Viscoelastic Deadly Fluid in Carnivorous Pitcher Plants. PLoS ONE 2(11): e1185 doi:10.1371/journal.pone.0001185

fluid to a stress depends on both the viscosity and the density of the fluid. Both the inertia and the viscosity of a fluid affect how quickly flows will start and stop. So it is often useful to talk about the kinematic viscosity of a fluid:

$$\text{Kinematic viscosity} = \nu = \frac{\eta}{\rho}$$

This kinematic viscosity, in a sense, compares the ability of friction in the fluid to stop the flow to the inertia of the fluid. A fluid with a lot of friction will quickly damp out any flows which begin. A fluid with high density will produce flows which continue for a long time. The symbol ν is often used when referring to kinematic viscosity.

A good example is provided if you imagine spinning a bucket of water or a bucket of air. If you do spin both for long enough to have the internal fluid spinning with the bucket and then suddenly stop the bucket itself. The fluid will continue to rotate for some time. A whirlpool of air stops spinning much faster than a whirlpool of water, because although the viscosity of air is 50 times less than that of water, the density is 1000 times less. So although there is somewhat less viscous force available to stop the air, much less is needed, and it stops much more quickly. The difference in kinematic viscosity of these two expresses this different response to change much better than a simple comparison of their viscosities.

18.4 Real fluid flow and turbulence

We have been discussing fluid flow which is smooth and laminar, neatly layered and without mixing. But that's not what we often see in fluid flow. Instead we see a swirling mixing of fluids, a much more complicated flow in which stability is never evident. What causes this turbulence, and what marks the transition between smooth flow and turbulence?

When the motion of a fluid is dominated by internal friction, by viscous drag, the flow will always be smooth. Any deviations from smoothness will be "damped out" by the friction before they have a chance to become large. This is what you usually see when you pour highly viscous liquids like honey or syrup. Just try mixing a jar of honey with a spoon and you'll see what I mean.

On the other hand, when fluid friction is small and something disturbs the flow even slightly, bits of the fluid which are knocked out of place will be able to travel far from where they would be in smooth flow before viscous effects stop them. This allows them to affect other parts of the fluid before they are stopped, leading to a kind of cascade of confusion. This is what you usually see in the flow of air, perhaps most clearly when you look at the smoke from a recently extinguished candle. Swirling around, it doesn't take long for the smoke to be well mixed with the air around it.

Reynold's number as the thing which characterizes fluid flow

Somehow the important thing is the balance between the inertia of a bit of the fluid, and the size of the viscous forces acting on it. What follows is a generic discussion of how these two things might vary with the parameters of the material.

First consider the inertia of the object: what size force does it take to stop it? The required force is given by its mass multiplied by some desired acceleration. Lets imagine w little cube of fluid with edge length L , and ask that we try to stop it before it travels through a distance equal to its size.

$$F = ma = (\rho L^3) a = (\rho L^3) \frac{v}{\Delta t} = (\rho L^3) \frac{v}{L/v} = \rho L^2 v^2$$

Where we have assumed $\Delta t = L / v$ is a reasonable approximation for how quickly we might make it stop. That is, I want it to stop before it would travel its own length, if it continued at its initial speed v . So this expression gives us an estimate of how large a force is required to stop this bit of fluid when it tries to move relative to the flow.

Now we examine the available viscous forces: how large a force is available to stop it?

$$\frac{F}{A} = \eta \frac{v}{L} \quad \text{implies} \quad F = \eta \frac{v}{L} L^2 = \eta v L$$

The ratio of these two forces is a dimensionless number:

$$R = \frac{\text{force required to stop}}{\text{force available from viscosity}} = \frac{\rho L^2 v^2}{\eta v L} = \frac{\rho L v}{\eta}$$

This is the famous Reynolds' number R , named for British fluid mechanic Osborne Reynolds. Note that it is dimensionless, a pure number independent of what units we measure it in. What does this dimensionless nature mean? It suggests that R is an important number, because it allows us to compare problems on all kinds of scales, independent of the units we use, so that they can be uniformly understood.

Meaning of Reynolds' number

R is dimensionless, so it only has a value, not units. Remember that it's the ratio of inertial properties to viscous forces. When R is small, then the viscous forces are dominant and the resulting flow will be smooth and orderly. When R is large, viscous forces are negligible and the resulting flow rapidly becomes complex and turbulent.

What is "large" and "small"? We can guess from how we formulated it. When $R \ll 1$, the available friction force is much larger than required to stop a bit of the fluid from getting out of place. When $R = 1$, the available friction force is able to stop the bit of fluid, but only after it travels a distance about its own size. This would surely lead to some mixing. When $R \gg 1$, the available viscous forces will be unable to prevent mixing on quite large scales, and the flow will become thoroughly turbulent.

While it is useless to define precise boundaries in Reynold's number which characterize different kinds of flow, one might fruitfully say that when $R < 10^{-3}$, flows are completely dominated by viscosity, when $R > 3000$ flows are fully turbulent, and when $10^{-3} < R < 3000$, the situation is less clear, with some mixing but still substantial layering and order in the flow.

We brushed over a problem though, what is this length scale L we have used in determining R ? L is the thing which sets the scale for the problem. Could be the size of an object moving through a fluid, diameter of a pipe through which it flows, etc. It's the scale on which we want to examine the fluid.

We can even ask about different length scales within the same problem. Imagine the flow of water in a pipe of radius r_{pipe} . If we wish to know about the flow in the pipe overall, we might use this length scale. But imagine that we want to know what the flow is like around little bumps on the wall. If these are smaller, with size $r_{\text{bump}} \ll r_{\text{pipe}}$, the flow around them may have a different nature. For example, the Reynolds' number for the pipe as a whole might be large (both r_{pipe} and v are large for the whole pipe) while the Reynold's number for the bumps may be small (both r_{bump} and $v_{\text{near wall}}$ are small). This is very typical of flows; turbulence on large scales accompanied by smooth laminar flow on small scales.

Implications of scale dependence

The scale dependence (the dependence on L) in the Reynolds' number shows that what we will see depends on what we look at. Let's consider some large and small things.

Consider a falling dust particle. Here the density is the density of the air through which the particle falls, $\rho_{\text{air}} = 1.3 \text{ kg/m}^3$. The velocity of such a falling dust particle is small, let's guess one millimeter per second. The size of a typical dust grain might be 10^{-4} m , and we know that the viscosity of the air is $\eta = 1.8 \times 10^{-5} \text{ Ns/m}^2$. Putting these all together, we get a Reynold's number:

$$R = \frac{(1.3 \text{ kg/m}^3)(10^{-3} \text{ m/s})(10^{-4} \text{ m})}{(1.8 \times 10^{-5} \text{ Ns/m}^2)} = 0.007$$

This means a falling dust grain is pretty well in the viscous regime and will always experience the small-slow kind of friction.

Redoing this for a falling raindrop is instructive. The fluid properties remain the same, but now the terminal velocity is larger, and the length scale is bigger as well. Let's assume a velocity of 0.1 m/s and a length scale of a half centimeter. Putting these together we find:

$$R = \frac{(1.3 \text{ kg/m}^3)(0.1 \text{ m/s})(5 \times 10^{-3} \text{ m})}{(1.8 \times 10^{-5} \text{ Ns/m}^2)} = 36$$

Here the motion is dominated by the inertia of the fluid, but affected still by the friction which is required to slip one layer of fluid over another. Calculating this once more for a falling person, where the terminal velocity might be 30 m/s and the length scale two meters, we get:

$$R = \frac{(1.3 \text{ kg/m}^3)(30 \text{ m/s})(2 \text{ m})}{(1.8 \times 10^{-5} \text{ Ns/m}^2)} = 4.3 \times 10^6$$

Here the motion is completely dominated by the inertia of the fluid. The fact that there is viscosity at all, that there is fluid friction, is really not important.

Reynolds' number describes the quality of a flow, giving an idea of its essential nature. If you want to know whether a flow is laminar or turbulent, start by calculating the Reynolds' number

So if you want to model a flow, by placing a model in a wind tunnel for example, you have to do it with the same R as you would have in the real system. This means that if you use a $1/10$ scale model (L becomes $L/10$) you have to somehow compensate for this in your test, perhaps by going $10x$ faster. This presents real problems for the testing of aircraft, as going $10x$ faster would usually require traveling faster than sound. Other problems occur when you try to blast air over a model at speed greater than the speed of sound.

18.5 Life at large and small Reynolds' number

People live in a world where the Reynolds' number is almost always large. We move with large velocities, live in air (which has a relatively small viscosity) and are ourselves large. So are motions through the air are usually not dominated by viscosity. Perhaps more important, we also live in a world where turbulent mixing is easy to achieve. If we want to mix cream into our coffee uniformly we just do it. With no special effort we can make the uniformly smooth liquid we so enjoy. The air around us is continually mixing on the scales we are about. When you breath out some air enriched in CO_2 , you can be confident that the next breath you take in will already be mixed so that new oxygen will have entered and old CO_2 gone out.

Bacteria, on the other hand, live in a world totally dominated by viscosity. Their speed v is small, and size is small, and they live mostly in water. So they're always in the viscous dominated, low Reynolds' number regime. Every time such a creature stops pushing for forward motion, it immediately stops moving. This is a world in which $F \propto v$, rather than $F \propto a$. It's an Aristotelian world in which you have to have a forward force constantly to have motion. As soon as it disappears you stop.

In addition, the fluid in the world of low Reynolds' number does not mix. Since it flows only laminarily, it never gets the kind of turbulent mixing that allows cream to mix into your coffee. This is very important if you're a bacteria hoping that a little food will come your way; basically it won't.

18.6 Reynold's number and terminal velocity

Now at last we can say what we mean by small-slow and large-fast when we talk about what form of fluid friction is likely to affect an object. When the Reynolds' number for an object moving through a fluid is low, the small-slow form of friction will apply. When the Reynolds' number is large, the large-fast form of friction will apply. At intermediate Reynolds' number, neither of these two forms for fluid friction will give an accurate description of the motion.

Remember when we calculated the motion of an object falling through a fluid. We found that the small-slow friction case had a nice analytic solution, while the large-fast friction case did not. Now you can see that most of the objects you care about will actually be in the large-fast regime. Only when you're considering tiny things and microscopic flows will you have small-slow friction.

A Quick Summary of Some Important Relations

Fluid flow and continuity:

For incompressible fluids, flow rates at different points in the flow can be related:

$$A_1 v_1 = A_2 v_2$$

Energy conservation in ideal fluid flows:

In a steady flow, we expect energy to be spread equally throughout the flow. Bernoulli's equation accounts for the fact that this energy may take different forms at different locations in the flow.

$$P_1 + \rho g h_1 + \frac{1}{2} \rho v_1^2 = P_2 + \rho g h_2 + \frac{1}{2} \rho v_2^2$$

This equation shows how pressure in one place can be generated by greater height in another, pressure can be converted to kinetic energy, kinetic energy can be converted to height, etc. Quantitative use of this principle is limited - it leaves out energy losses due to friction – but qualitatively it is very helpful.

Viscosity and friction in a fluid:

Friction in the flow of a fluid can often be expressed as viscosity, defined by this relation:

$$\sigma_{\text{shear}} = \frac{F}{A} = \eta \left(\frac{dv}{dy} \right)$$

Shear flow rate, the increase in velocity from one layer to the next, is determined by shear stress and viscosity. When viscosity η is a constant, we say the fluid is 'Newtonian'.

Flow of a viscous fluid in a pipe:

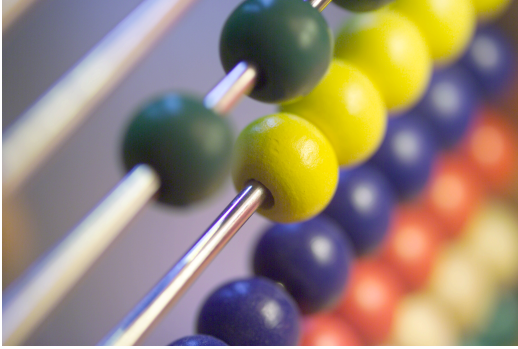
Flow of a Newtonian fluid in a pipe can be modeled as Poiseuille flow.

$$v_{\text{max}} = \frac{r_{\text{pipe}}^2}{4\eta} \frac{\Delta P}{L_{\text{pipe}}}$$
$$\Phi = \frac{\pi r_{\text{pipe}}^4}{8\eta} \frac{\Delta P}{L_{\text{pipe}}}$$

Reynold's number characterize quality of flow:

Flow may be either layered and smooth, dominated by friction, or turbulent and mixed, dominated by inertia. Reynold's number characterizes the nature of the flow on a particular scale.

$$R_e = \frac{\rho L v}{\eta}$$



In a Nutshell

“Mathematics is a language plus reasoning; it is like a language plus logic. Mathematics is a tool for reasoning.”

— Richard P. Feynman (The Character of Physical Law)

Physics frequently makes use of mathematics to describe natural phenomena. Most laws in physics are expressed in mathematical form, and new laws can be found by mathematically combining known ones. Consequently, most mathematical methods were first developed in the context of physics.

It is said that “a picture is a thousand words” — that’s why in physics we often use graphs and sketches of situations. But, like in the Feynman quote above, a formula is also like a thousand words; if understood correctly as a language, a formula says how things are related to each other, and even more: mathematical manipulations can lead to new insights, new discoveries. Why our universe functions that way is a mystery, but it makes our job as physicists easier: we don’t have to memorize so much unconnected garbage. To quote Feynman again: “In fact the total amount that a physicist knows is very little. He has only to remember the rules to get him from one place to another and he is all right” In this course, we will frequently use mathematical reasoning to derive how to get from one place to the other.

Some students ignore derivations, thinking that instructors and textbook authors merely list them to prove that the final result is correct. Instead, those students ask for a handy summary of all the “useful” formulas that one can use to plug numbers into and get new numbers out of. Wrong.

Unfortunately, when studying physics, mathematics oftentimes becomes an additional hurdle. In this chapter, we will go over most of the mathematical toolset needed for the remainder of this course. The main intention is to make you aware of what lies ahead, and to allow additional time to prepare yourself if needed. The chapter is written in concise form since the majority of the material has been dealt with in prerequisite courses already. If concepts are unfamiliar, please refer to external sources (textbooks, etc) to re-familiarize yourself with those ideas.

The good news: if you master this chapter, there should be no surprises anymore in this course as far as mathematics goes — so you can concentrate on the physics surprises.

1.1 Functional Dependencies

In physics, one often needs to express that one quantity changes depending on another one. For example, the fact that a changes when t changes would be expressed by writing

$$a(t)$$

or “ a is a function of t .” Another example would be

$$F(x, y)$$

meaning, F is a function of both x and y . It’s nice for the reader to have a complete listing of all variables that a function depends on, but sometimes physicists are lazy. Functional values are then obtained simply by plugging in those variables.

Example 1.1 A Function

Given

$$F(x, y) = 3(x^2 + y^2)$$

for $x = 3$ and $y = 4$, it is $F(3, 4) = 75$.

Unfortunately, there is no good notation to express the fact that a function does *not* depend on a certain variable, i.e., that if you change that variable, nothing happens. For example, $F(x, y) = 3(x^2 + y^2)$ does not depend on z . You might hear, though, that “ F is independent of z ,” which means just that.

Sometimes, the exact form of a function is less important than its general functional form or class. You might simply want to know if a function gets bigger of smaller when something else changes. Important classes of functional dependence are:

Linear The value increases proportionally, e.g.,

$$f(t) = vt$$

One might say: “ f increases linearly with t ” or “ f is proportional to t ” — if nothing is said, we mean *directly* proportional.

Quadratic The value increases proportionally to the the variable squared, e.g.,

$$f(t) = at^2$$

One might say: “ f increases quadratically with t ”

Exponential The value increases exponentially, e.g.,

$$f(t) = e^{kt} = \exp(kt)$$

or

$$f(t) = a^t$$

Inversely proportional The value decreases with one over the value, e.g.,

$$f(t) = \frac{a}{t}$$

One might say: “ f is inversely proportional to t ,” or “ f falls off linearly with t .” Some people also say “indirectly proportional” (again, if nothing is said, we mean “directly”).

Inversely quadratic The value decreases with one over the value squared, e.g.,

$$f(t) = \frac{a}{t^2}$$

One might say: “ f falls off quadratically with t .”

1.2 Quadratic Equations

A **quadratic equation** such as

$$ax^2 + bx + c = 0$$

has two solutions

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

An alternative form is

$$x^2 + px + q = 0$$

with $p = b/a$, $q = c/a$, and the solutions

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}$$

Of course, the argument of the square root could become negative, in which case you get imaginary solutions — that’s okay, often that will be where the interesting physics starts.

Example 1.2 A Quadratic Equation

The larger of the two solutions for

$$-2x^2 + 19x + 12 = 0$$

is approximately 10.0944.

1.3 Trigonometry

1.3.1 Degrees and Radians

In physics, we frequently use two different ways to specify angles: in degrees or in radians. Both ways are valid, and which one is chosen often depends only on convenience or on what is more intuitive in a given situation.

In degrees, a full circle is 360° , in radians, it is 2π . You can thus convert back and forth through

$$\begin{aligned}(\text{angle in degrees}) &= \frac{180}{\pi}(\text{angle in radians}) \\(\text{angle in radians}) &= \frac{\pi}{180}(\text{angle in degrees})\end{aligned}$$

For frequently used angles, it is

Degrees	Radians
0°	0
45°	$\pi/4$
90°	$\pi/2$
180°	π
360°	2π

When using a pocket calculator, you need to specify which angle measure you are using. There are usually keys for **RAD** or **DEG**, or **MODE** to toggle back and forth. Not having selected the correct mode is a common source of error — again, either mode is fine, you just need to stick with it throughout the problem.

The only place where you have to use radians and not degrees is the so-called small-angle approximation $\sin(x) \approx x$, which is only correct if x is given in radians.

Also, after doing any kind of calculus (derivatives, etc), keeping track of degrees when plugging in values becomes extremely painful very quickly and should better be avoided — use radians!

1.3.2 Trigonometric Functions

The trigonometric functions sine, cosine, and tangent are plotted versus degrees in Figure 1.1. All of these functions are periodic with a period of 2π (or 360°).

It often comes in handy to know just a few values by heart:

$$\begin{aligned}\sin(0) &= 0 & \cos(0) &= 1 \\ \sin(30^\circ) &= 0.5 & \cos(60^\circ) &= 0.5 \\ \sin(90^\circ) &= 1 & \cos(90^\circ) &= 0 \\ \tan(0) &= 0 \\ \tan(90^\circ) &= \infty\end{aligned}$$

The inverse of the trigonometric functions are the arc-functions, e.g., \arcsin , \arccos , etc. Since the trigonometric functions are periodic (for example, $\sin(185^\circ) = \sin(-5^\circ) = \sin(545^\circ)$), it is not necessarily $\varphi = \arcsin(\sin(\varphi))$ — the arc-function results are only what you expect up to a phase, i.e., an integer multiple of 2π .

1.3.3 Triangles

Trigonometric functions can be used to calculate the lengths of the sides of a triangle with one right angle (90° (degrees) or $\pi/2$ (radians)). For example, in the triangle in 1.2, it is

$$\begin{aligned}\sin \theta &= b/a \\ \cos \theta &= c/a \\ \tan \theta &= b/c\end{aligned}$$

1.3.4 Trigonometric Identities

Two handy so-called trigonometric identities are:

$$\sin^2 \theta + \cos^2 \theta = 1 \quad \sin(2\theta) = 2 \sin \theta \cos \theta$$

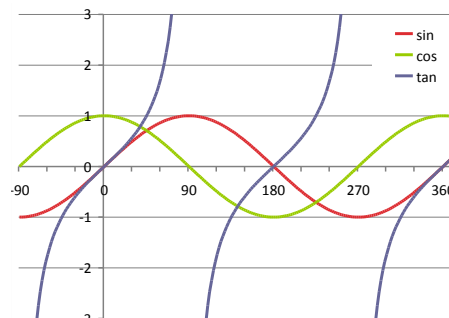


Figure 1.1: *The trigonometric functions*

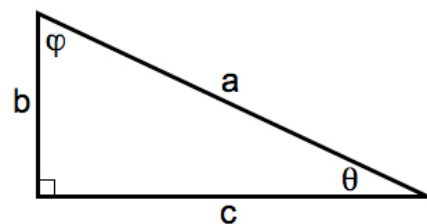


Figure 1.2: *A right-angle triangle*

Some useful angle relations are:

$$\begin{aligned}\sin(-\theta) &= -\sin(\theta) & \cos(-\theta) &= \cos(\theta) \\ \sin(90^\circ - \theta) &= \cos(\theta) & \cos(90^\circ - \theta) &= \sin(\theta) \\ \sin(180^\circ - \theta) &= \sin(\theta) & \cos(180^\circ - \theta) &= -\cos(\theta) \\ \sin(270^\circ + \theta) &= -\cos(\theta) & \cos(270^\circ + \theta) &= \sin(\theta) \\ \sin(\alpha + \beta) &= \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)\end{aligned}$$

1.4 Exponential Function

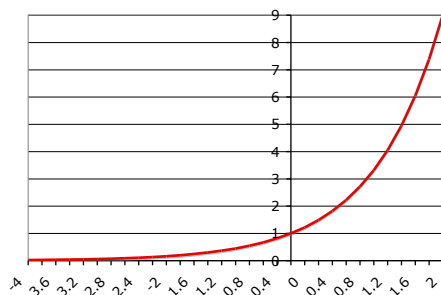


Figure 1.3: *The exponential function*

The exponential function

$$\exp(x) = e^x$$

with the Euler number e is plotted in Figure 1.3. Physicists often use the $\exp(x)$ rather than the e^x notation, since x might become a rather large expression, and it looks awkward to have a little e with a huge exponent, e.g.,

$$e^{i \cdot \arctan\left(\frac{b\Omega}{m(\omega^2 - \Omega^2)}\right)}$$

just looks rather confusing, compared to

$$\exp\left(i \cdot \arctan\left(\frac{b\Omega}{m(\omega^2 - \Omega^2)}\right)\right).$$

If $f(x) = \exp(kx)$ (with some random positive k), one would say that “ f increases exponentially with x .” If $f(x) = \exp(-kx)$ (again with some random positive k), one would say that “ f falls off exponentially with x .” Another common situation is $f(x) = a(1 - \exp(-kx))$ — for small x , this function would be zero, while for big x , it would be a .

The inverse of the exponential function is the \ln -function, the natural logarithm. It is $x = \ln(\exp(x))$ and $x = \exp(\ln(x))$. Additional rules for the logarithm and exponential functions are:

$$\begin{aligned}x^y &= \exp(y \cdot \ln(x)) \\ \exp(x)\exp(y) &= \exp(x + y) & \ln(x \cdot y) &= \ln(x) + \ln(y) \\ \frac{\exp(x)}{\exp(y)} &= \exp(x - y) & \ln\left(\frac{x}{y}\right) &= \ln(x) - \ln(y)\end{aligned}$$

1.5 Calculus

1.5.1 Differentials

Physicists usually use the “ d/dx ”-notation to denote differentials. Basically, it is

$$dx = \lim_{\Delta x \rightarrow 0} \Delta x,$$

where Δx is some finite difference of x -values. The derivative of a function $f(x)$ at a location a with respect to x is then defined as

$$\frac{df}{dx}(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} = \lim_{\Delta x \rightarrow 0} \frac{f(a + \Delta x) - f(a)}{\Delta x} = \frac{f(a + dx) - f(a)}{dx}.$$

Differentiation is also often written as an operator,

$$\frac{d}{dx}f(x) \quad \text{or} \quad \frac{d}{dx}(\dots)$$

where the d/dx differentiates any “factor” behind it.

1.5.2 Derivative of a Polynomial

The derivative of

$$ax^n$$

with respect to x is

$$\frac{d}{dx}ax^n = nax^{n-1}$$

Since derivatives are additive, i.e.,

$$\frac{d}{dx}(f(x) + g(x)) = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$$

the derivative of for example

$$ax^n + bx^m$$

is

$$\frac{d}{dx}(ax^n + bx^m) = \frac{d}{dx}ax^n + \frac{d}{dx}bx^m = nax^{n-1} + mbx^{m-1}$$

Example 1.3 Derivatives of Polynomials

$$\frac{d}{dx}(3ax^5 + 2a^2x^4 + a^6) = 15ax^4 + 8a^2x^3$$

$$\frac{d}{da}(3ax^5 + 2a^2x^4 + a^6) = 3x^5 + 4ax^4 + 6a^5$$

$$\frac{d}{dy}(3ax^5 + 2a^2x^4 + a^6) = 0$$

1.5.3 Derivative of Sine, Cosine, and Exponentials

It is

$$\frac{d}{dx}a \sin(bx) = ab \cos(bx)$$

$$\frac{d}{dx}a \cos(bx) = -ab \sin(bx)$$

$$\frac{d}{dx}a \exp(bx) = ab \exp(bx)$$

The above shows why these functions are so important for physics: you take the derivative of the exponential function, and you get it back. You take the second derivative of sine and cosine, and you get them back.

Example 1.4 Derivatives of Sine, Cosine, and Exponentials

$$\frac{d}{dt}A \sin(\omega t) = A\omega \cos(\omega t)$$

$$\frac{d}{dx}L \exp(-kx) = -kL \exp(-kx)$$

$$\frac{d}{dx}A \cos(\omega t) = 0$$

1.5.4 Rules for Derivatives

The following rules help in calculating derivatives:

$$\frac{d}{dx}((f(x)g(x))) = g(x)\frac{d}{dx}f(x) + f(x)\frac{d}{dx}g(x) \quad (\text{product rule})$$

$$\frac{d}{dx}f(g(x)) = \frac{df(g)}{dg} \frac{dg(x)}{dx} \quad (\text{chain rule})$$

$$\frac{d}{dx}(cf(x)) = c\frac{d}{dx}f(x) \quad (\text{special case of chain rule})$$

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{\frac{df(x)}{dx}g(x) - f(x)\frac{dg(x)}{dx}}{g^2(x)} \quad (\text{quotient rule})$$

Example 1.5 There is More Than One Way

There is more than one way to calculate the the derivative of $\sin^2 x = (\sin x)^2$ with respect to x :

- you could use the product rule:

$$\begin{aligned}\frac{d}{dx} \sin^2 x &= \frac{d}{dx} \sin x \sin x \\ &= \frac{d \sin x}{dx} \sin x + \sin x \frac{d \sin x}{dx} \\ &= \cos x \sin x + \sin x \cos x \\ &= 2 \sin x \cos x\end{aligned}$$

- you could also use the chain rule, with $g(x) = \sin x$ and $f(g) = g^2$:

$$\begin{aligned}\frac{d}{dx} \sin^2 x &= \frac{d}{dx} (\sin x)^2 \\ &= \frac{d(\sin x)^2}{d \sin x} \cdot \frac{d \sin x}{dx} \\ &= 2 \sin x \cdot \cos x\end{aligned}$$

1.5.5 Series Expansions

If a function can be differentiated over and over, it can be expanded in a series around a point a (d^n/dx^n is the n th derivative; the zeroth derivative is the function itself):

$$f_a(x) = \sum_{n=0}^{\infty} \frac{d^n f(a)}{dx^n} \frac{(x-a)^n}{n!} = f(a) + \frac{df(a)}{dx}(x-a) + \frac{d^2 f(a)}{dx^2} \frac{(x-a)^2}{2} + \frac{d^3 f(a)}{dx^3} \frac{(x-a)^3}{6} + \dots$$

For well-behaved functions, the series converges quickly if $|x-a| \ll 1$, so already after a few terms gives a decent approximation of the function itself around a .

The following are some useful series expansions around $a = 0$, which can be used to approximate a function for $|x| \ll 1$:

$$\begin{aligned}(1+x)^k &= 1 + kx + \frac{k(k-1)}{2}x^2 + \dots \\ \sqrt{1+x} &= 1 + \frac{x}{2} + \dots \\ \frac{1}{\sqrt{1-x}} &= 1 + \frac{x}{2} + \dots \\ \exp(x) &= \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \dots \\ \sin(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{6} + \dots \quad (x \text{ in radians}) \\ \cos(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \dots \quad (x \text{ in radians}) \\ \tan(x) &= x + \frac{x^3}{3} + \frac{2x^5}{15} + \dots \quad (x \text{ in radians, } |x| < \frac{\pi}{2})\end{aligned}$$

Sample Problem 1.1 Approximation

Q: How does the function

$$Q(t) = Q_0 (1 - \exp(-\beta t))$$

behave for small values of t ?

A:

$$Q(t) = Q_0 (1 - (1 - \beta t + \dots)) \approx Q_0 \beta t$$

For small t , $Q(t)$ increases linearly from zero.

1.5.6 Differential Equations

Differential equations are those relating a function to its own derivative(s). They are frequently solved by “trial and error,” where the trial function is called “ansatz.”

Example 1.6 Differential Equation

Differential equation:

$$\frac{d}{dt}f(t) = kf(t)$$

We are looking for a function $f(t)$, where the first derivative is just some factor times the function itself.

- Ansatz:

$$f(t) = at^2 + bt + c$$

Calculate first derivative:

$$\frac{d}{dt}f(t) = 2at + b$$

Insert into differential equation:

$$2at + b = k(at^2 + bt + c)$$

This ansatz does not work, since no non-zero a , b , and c can be found such that this equation is true for all t .

- New ansatz:

$$f(t) = a \exp(bt)$$

First derivative:

$$\frac{d}{dt}f(t) = ab \exp(bt)$$

Insert into differential equation:

$$ab \exp(bt) = ka \exp(bt)$$

The ansatz works for $b = k$ and any value of a . In a physics problem, a would need to be found from so-called boundary conditions, for example, from the value the function needs to have for $t = 0$.

1.5.7 Antiderivatives and Integrals

Antiderivatives are the reverse of derivatives. I.e., if

$$f(t) = \frac{d}{dt}g(t)$$

then $f(t)$ is the derivative of $g(t)$, and $g(t)$ is the antiderivative of $f(t)$, both with respect to t .

Antiderivates are only determined up to a constant (i.e., a term that does not depend on the variable you do the derivative with respect to). That is because the derivative of such a term would be zero.

Example 1.7 Antiderivative

$$f(t) = at$$

An antiderivative of $f(t)$ would be

$$g(t) = \frac{1}{2}at^2$$

but also the function

$$h(t) = \frac{1}{2}at^2 + c$$

since the derivative of the constant c with respect to t is zero.

Since

$$f(t) = \frac{dg(t)}{dt} = \frac{g(t+dt) - g(t)}{dt} \implies g(t+dt) = g(t) + f(t)dt$$

we can get

$$g(t_b) = g(t_a) + \int_{t_a}^{t_b} f(t)dt$$

or

$$\int_{t_a}^{t_b} dt f(t) = g(t)|_{t_a}^{t_b} = g(t_b) - g(t_a)$$

(the “factor” dt can be put anywhere, but it is usually stuck in either directly after the integral sign or at the very end). Graphically, integrals can be determined as the area under the graph of a function.

Integrals can be used to calculate the average of a function over an interval:

$$f_{\text{ave}} = \frac{1}{t_b - t_a} \int_{t_a}^{t_b} dt f(t)$$

1.5.8 Partial Derivatives

The concept of **partial derivatives** is useful when a function depends on a number of variables, for example, some

$$f(a, b, c)$$

A partial derivative with respect to for example b is defined as

$$\frac{\partial f(a, b, c)}{\partial b} = \lim_{\Delta b \rightarrow 0} \frac{f(a, b + \Delta b, c) - f(a, b, c)}{\Delta b} \quad a, c = \text{constant}$$

You may well ask how this is any different from the “normal” derivative, and in fact, it usually isn’t.

The difference kicks in if somehow for example a depends on b , i.e., $a(b)$, due to some other external circumstance. Let’s say

$$f(a, b, c) = 2a^2 + 3b^2 + 4c^3$$

and for example for some odd reason, it always has to be

$$a(b) = 3b^2 \quad c(b) = 3b^4$$

Now the partial derivative completely ignores the latter:

$$\frac{\partial}{\partial b} f(a, b, c) = 6b$$

— this is what make it “partial” — you could also say “lazy.”

Now, if you want to get the **total derivative**, you would have to do

$$\begin{aligned} \frac{d}{db} f(a, b, c) &= \frac{\partial f(a, b, c)}{\partial b} + \frac{\partial f(a, b, c)}{\partial a} \frac{da}{db} + \frac{\partial f(a, b, c)}{\partial c} \frac{dc}{db} = 6b + 4a \cdot 6b + 12c^2 \cdot 12b^3 \\ &= 6b + 4(3b^2) \cdot 6b + 12(3b^4)^2 \cdot 12b^3 = 6b + 72b^3 + 1296b^{11} \end{aligned}$$

Of course, you would have to get the same thing if you first plugged $a(b) = 3b^2$ and $c(b) = 3b^4$ into f , i.e., $f(3b^2, b, 3b^4)$:

$$\frac{d}{db} f(3b^2, b, 3b^4) = \frac{d}{db} (2(3b^2)^2 + 3b^2 + 4(3b^4)^3) = \frac{d}{db} (18b^4 + 3b^2 + 108b^{12}) = 72b^3 + 6b + 1296b^{11}$$

Amazing!

Oftentimes, the following abbreviation is used:

$$\partial_a = \frac{\partial}{\partial a} \quad \partial_a^n = \frac{\partial^n}{\partial a^n}$$

In summary, don’t panic when you see a partial derivative: ∂ is easier than d .

1.6 Complex Numbers

At the center of complex numbers is the imaginary number i , where

$$i = \sqrt{-1}$$

Engineers often call this number j , since they use i for currents.

A complex number z then has two components, the real component a and the imaginary component b (a and b both being real (non-complex) numbers) with

$$z = a + ib$$

Rules for algebra with complex numbers follow from the property of i . With $y = c + id$ and $z = a + ib$ it is:

$$\begin{aligned} y + z &= (c + id) + (a + ib) = (c + a) + i(d + b) \\ y \cdot z &= (c + id) \cdot (a + ib) = (ac - bd) + i(bc + ad) \end{aligned}$$

In physics, complex numbers are frequently used in connection with the exponential function, where they connect the exponential and trigonometric functions with each other:

$$\exp(i bt) = \cos(bt) + i \sin(bt)$$

or, more general,

$$\exp(z t) = \exp((a + ib)t) = \exp(at) \exp(i bt) = \exp(at) (\cos(bt) + i \sin(bt))$$

For example:

$$(a + ib) \exp(i ct) = ((a - b) + i(a + b)) (\cos(ct) + i \sin(ct))$$

The magnitude of a complex number

$$z = x + iy$$

is

$$|z| = \sqrt{x^2 + y^2}$$

and every complex number can be written as

$$z = |z| \cdot \exp(i \theta) \quad \text{with} \quad \theta = \arctan\left(\frac{y}{x}\right)$$

where θ is called the phase.

Sample Problem 1.2 Magnitude

Q: What is the magnitude of $\exp(i \omega t)$?

A:

$$|\exp(i \omega t)| = \sqrt{\cos^2(\omega t) + \sin^2(\omega t)} = 1$$

1.7 Vectors and Matrices

1.7.1 Vectors and Scalars

- For our purposes, **scalars** are just single numbers, e.g., 17 or 42. Scalar quantities can have units attached, such as “40 meters” or “4 kilograms.” An important property of scalars is that, as opposed to vectors, they do not depend on any kind of coordinate system — scalars have no directions.
- On the other hand, **vectors** denote directions. Their mathematical representation is usually a “bundle of numbers,” each of which is called a component or coordinate (which can have units). In a wider sense, the statement “drive 10 miles east, and another 2 miles south” is a vector (see Figure 1.4, with the two components (or coordinates) “10 miles east” and “2 miles south”).

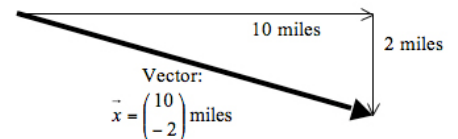


Figure 1.4: A vector

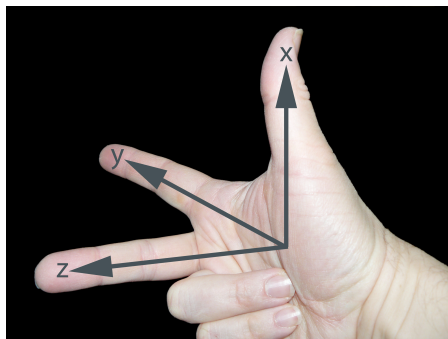


Figure 1.5: *Right-handed orientation of axes*

Vector representations only make sense if the directions are defined ahead of time, such as “east” and “south” — you need to know where “south” is, i.e., the **coordinate system**. In physics, we usually use the so-called **cartesian coordinate system**, with its three directions (often called “dimensions”) left-right, forward-backward, and up-down, referred to as x , y , and z dimensions. These directions are oriented according to the **right-hand rule**, see Figure 1.5.

Vectors are often written with a little arrow above them to denote their directional character, e.g.

$$\vec{a}$$

The components of the vector are usually written as the name of the vector with the subscripts x , y , and z , so for example, a_y would be the y -component of the vector \vec{a} and corresponds to how much the vector \vec{a} points forward.

What in actuality is “forward” of course depends on the point of view, or the orientation of the coordinate system. In this example, to describe “backward,” we use negative numbers — next time, when somebody asks for directions, tell them to drive minus ten miles north if you want them to go south. Also, location coordinates depend on the zero-point (so called origin) of the coordinate system, which we need to define. Where we put the origin and how we orient the axes is up to us, and usually a matter of convenience: there are more and less *convenient* ways where to point the axes, but there are no *wrong* ways; we will explore the consequences of this in Chapter 4 (page 63).

If one wants to write down the complete vector, the components are written in brackets, e.g.,

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \quad \text{or} \quad \vec{a} = (a_x, a_y, a_z)$$

In this course, in line with common practice, we use the two notations interchangeably, depending simply on layout considerations. In a Linear Algebra course or in graduate level physics courses, the two notations are used to express different things. Units are usually written after the vector, for example

$$\vec{p} = \begin{pmatrix} 17 \\ 42 \\ 0 \end{pmatrix} \frac{\text{kg} \cdot \text{m}}{\text{s}}.$$

Example 1.8 Vectors

To express that a car is moving with 10 miles per hour to the left, you might say that its “velocity vector” \vec{v} is

$$\vec{v} = \begin{pmatrix} -10 \\ 0 \\ 0 \end{pmatrix} \text{mph}$$

To express that something is located 10 meters in front of you, two meters to the right, and three meters underground, you could say that its “position vector” \vec{r} is

$$\vec{r} = \begin{pmatrix} 2 \\ 10 \\ -3 \end{pmatrix} \text{m}$$

When up and down are irrelevant, they are simply skipped, and we deal only with the two dimensions x and y . A vector

$$\vec{b} = \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$

would be a vector in “two dimensions,” i.e., nothing is happening in the third dimension and it is thus skipped.

Graphically, vectors are represented by arrows. When working in three dimensions, vectors pointing into the paper are represented by a circle with a cross (\otimes , you see the feathers of the arrow), and vectors pointing out of the paper are represented by circles with a dot in the middle (\odot , you see the tip of the arrow).

1.7.2 Vector-Valued Functions

In physics, vectors are usually not static, but depend on another variable, for example, time. This simply means that each component of the vector is a function, for example

$$\vec{a}(t) = \begin{pmatrix} a_x(t) \\ a_y(t) \\ a_z(t) \end{pmatrix}$$

Example 1.9 Vector-Valued Function

$$\vec{r}(t) = \begin{pmatrix} 3\frac{\text{m}}{\text{s}^2}t^2 \\ 4\frac{\text{m}}{\text{s}}t \\ 5\frac{\text{m}}{\text{s}^2}t^2 + 2\frac{\text{m}}{\text{s}}t \end{pmatrix}$$

For $t = 3\text{s}$, the vector is

$$\vec{r}(3\text{s}) = \begin{pmatrix} 27 \\ 12 \\ 51 \end{pmatrix} \text{m}$$

1.7.3 Vector Addition

Two vectors

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \quad \text{and} \quad \vec{b} = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix}$$

are added to a vector

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

simply by adding the components separately, i.e.,

$$\vec{c} = \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} + \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_x + b_x \\ a_y + b_y \\ a_z + b_z \end{pmatrix}$$

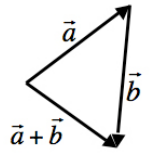


Figure 1.6: Vector Addition

Graphically, vectors are added by attaching the tail of the one vector to the head of the other, see Figure 1.6. To do $\vec{a} - \vec{b}$, you need to basically do $\vec{a} + (-\vec{b})$, where $-\vec{b}$ simply points in the opposite direction of \vec{b} . You could also think of this as attaching the head (instead of the tail) of \vec{b} to the head of \vec{a} .

Example 1.10 Vector Addition

$$\begin{pmatrix} 4 \\ 6 \\ 3 \end{pmatrix} + \begin{pmatrix} 2 \\ 9 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 15 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 4d + 2x \\ bx^2 \\ c \end{pmatrix} + \begin{pmatrix} 4x \\ 3x \\ e \end{pmatrix} = \begin{pmatrix} 4d + 6x \\ 3x + bx^2 \\ c + e \end{pmatrix}$$

1.7.4 Vector Length

The length (or magnitude) of a vector is calculated as

$$|\vec{a}| = \left| \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \right| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a^2} = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

A frequently used shorthand for the length of a vector is the same symbol without the arrow, e.g.,

$$\text{Shorthand: } a = |\vec{a}|$$

Example 1.11 Vector Length

$$\left| \begin{pmatrix} 3 \\ 2 \\ 5 \end{pmatrix} \right| = \sqrt{3^2 + 2^2 + 5^2} = \sqrt{38}$$

$$\left| \begin{pmatrix} c \\ 3b + c \\ 2a \end{pmatrix} \right| = \sqrt{c^2 + (3b + c)^2 + 4a^2}$$

1.7.5 Multiplying a Vector with a Scalar

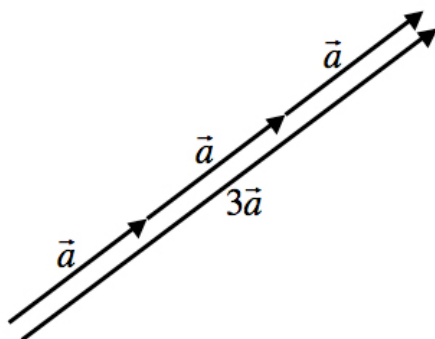


Figure 1.7: Multiplying \vec{a} by 3

A vector is multiplied with a scalar (number) through

$$c\vec{a} = c \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} = \begin{pmatrix} ca_x \\ ca_y \\ ca_z \end{pmatrix}$$

The new vector has the same direction (positive c) or the exact opposite direction (negative c), but its length changes by the factor c , i.e.,

$$|c\vec{a}| = c|\vec{a}|$$

Example 1.12 Multiplying a Vector with a Scalar

$$4 \begin{pmatrix} 3 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 4 \cdot 3 \\ 4 \cdot 2 \\ 4 \cdot 5 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \\ 20 \end{pmatrix}$$

$$b \begin{pmatrix} 3a \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 3ab \\ 2b \\ 5b \end{pmatrix}$$

The length of the vector changed by a factor b , i.e.,

$$\left| \begin{pmatrix} 3ab \\ 2b \\ 5b \end{pmatrix} \right| = \sqrt{(3ab)^2 + (2b)^2 + (5b)^2} = b\sqrt{(3a)^2 + 2^2 + 5^2} = b \left| \begin{pmatrix} 3a \\ 2 \\ 5 \end{pmatrix} \right|$$

1.7.6 Unit Vectors

A unit vector is simply a vector of length 1. It is frequently used if only a direction needs to be indicated. If you have a vector \vec{a} , a unit vector pointing in the direction of vector \vec{a} can be calculated as

$$\vec{e}_a = \frac{\vec{a}}{|\vec{a}|}$$

i.e., you divide the vector a by its own length. Unfortunately, several notations for unit vectors are in common use, e.g.,

$$\vec{e}_a \quad \text{or} \quad \vec{u}_a \quad \text{or} \quad \hat{a}$$

Example 1.13 Unit Vectors

The vector

$$\vec{a} = \begin{pmatrix} 4 \\ 6 \\ -2 \end{pmatrix}$$

has the unit vector

$$\vec{e}_a = \frac{1}{\sqrt{4^2 + 6^2 + (-2)^2}} \begin{pmatrix} 4 \\ 6 \\ -2 \end{pmatrix} = \frac{1}{\sqrt{56}} \begin{pmatrix} 4 \\ 6 \\ -2 \end{pmatrix} \approx \begin{pmatrix} 0.53 \\ 0.8 \\ -0.27 \end{pmatrix}$$

Frequently used unit vectors are the vectors along the coordinate axes:

$$\vec{e}_x = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \vec{e}_y = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \vec{e}_z = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and vectors are sometimes expressed in terms of them:

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} = a_x \vec{e}_x + a_y \vec{e}_y + a_z \vec{e}_z$$

Engineers seem to like the notation \hat{i} , \hat{j} , and \hat{k} for the same three coordinate unit vectors.

1.7.7 Polar Coordinates

In two dimensions, another way of specifying a vector is to give its length ρ and the angle θ it encloses with the positive x -axis. So you could say a vector is given by “length 5 meters, angle 30 degrees.”

Polar to Coordinate Representation: To get back to the coordinate representation, you would use

$$\vec{a} = \begin{pmatrix} \rho \cos \theta \\ \rho \sin \theta \end{pmatrix}$$

with ρ being the length (“5 meters”) and θ being the angle (“30 degrees”). ρ turns out to be the length a of the vector. As you look at the drawing, you can see that this is just straightforward trigonometry. That a is indeed the length of this vector, you can see from

$$|\vec{a}| = \sqrt{a^2 \sin^2 \theta + a^2 \cos^2 \theta} = a \sqrt{\sin^2 \theta + \cos^2 \theta} = a$$

using the trigonometric identity that $\sin^2 \theta + \cos^2 \theta = 1$ for any angle θ

Coordinate to Polar Representation: Given the cartesian coordinates,

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \end{pmatrix}$$

it is a bit more involved to get the polar representation, since the mechanism depends on the quadrant you are in:

- $a_x > 0; a_y > 0$: $\theta = \arctan(a_y/a_x)$
- $a_x < 0; a_y > 0$: $\theta = 90^\circ - \arctan(a_x/a_y)$
- $a_x < 0; a_y < 0$: $\theta = 180^\circ + \arctan(a_y/a_x)$
- $a_x > 0; a_y < 0$: $\theta = 270^\circ - \arctan(a_x/a_y)$

1.7.8 Cylindrical Coordinates

Cylindrical coordinates are one of two standard ways to extend polar coordinates into three dimensions. The angle θ remains the angle with the positive x -axis, and ρ is defined as the distance from the z -axis. In addition, you take the standard z -coordinate. If you translate these three coordinates back into cartesian coordinates, you get

$$\begin{pmatrix} \rho \cos \theta \\ \rho \sin \theta \\ z \end{pmatrix}$$

Note that here, ρ is *not* the length of the vector, instead, the length is $\sqrt{\rho^2 + z^2}$.

1.7.9 Spherical Coordinates

Another way of extending polar coordinates are spherical coordinates. The angle θ remains the angle with the positive x -axis, and in addition, the angle φ is the angle with the xy -plane, and a new coordinate r is the distance from the origin. Back into cartesian coordinates, it looks like this.

$$\vec{a} = \begin{pmatrix} r \cos \varphi \cos \theta \\ r \cos \varphi \sin \theta \\ r \sin \varphi \end{pmatrix}$$

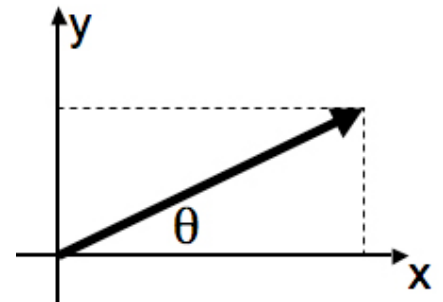


Figure 1.8: Polar Coordinates

In this case, r is the length a of the vector.

1.7.10 Scalar Derivative of a Vector

The scalar derivative of a vector is given by

$$\frac{d}{dt}\vec{a}(t) = \begin{pmatrix} \frac{d}{dt}a_x(t) \\ \frac{d}{dt}a_y(t) \\ \frac{d}{dt}a_z(t) \end{pmatrix}$$

i.e., the derivative of each component is calculated separately.

Example 1.14 Scalar Derivative of a Vector

$$\frac{d}{dt} \begin{pmatrix} 3at \\ 2t^2 \\ 5 \end{pmatrix} = \begin{pmatrix} 3a \\ 4t \\ 0 \end{pmatrix}$$

1.7.11 Scalar Product

The scalar product of two vectors is defined as

$$\vec{a} \cdot \vec{b} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \cdot \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = a_x b_x + a_y b_y + a_z b_z$$

Example 1.15 Scalar Product

$$\begin{pmatrix} 3 \\ 2 \\ 5 \end{pmatrix} \cdot \begin{pmatrix} 8 \\ 3 \\ 9 \end{pmatrix} = 3 \cdot 8 + 2 \cdot 3 + 5 \cdot 9 = 24 + 6 + 45 = 75$$

$$\begin{pmatrix} 3a \\ 2 \\ 5 \end{pmatrix} \cdot \begin{pmatrix} c \\ 3b+c \\ a \end{pmatrix} = 3ac + 2(3b+c) + 5a$$

The scalar product between two vectors is also equal to the product of the lengths of both vectors times the cosine of the angle θ between the vectors:

$$\begin{aligned} \vec{a} \cdot \vec{b} &= a_x b_x + a_y b_y + a_z b_z \\ &= |\vec{a}| |\vec{b}| \cos \theta \end{aligned}$$

Thus,

- if the two vectors are parallel ($\theta = 0$), the scalar product is just the product of the two lengths, because $\cos 0 = 1$
- if the two vectors are perpendicular ($\theta = 90^\circ$ in degrees, or $\theta = \pi/2$ in radians), the scalar product is zero, because $\cos 90^\circ = 0$

Example 1.16 Scalar Product and Angles

The vectors

$$\vec{a} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{b} = \begin{pmatrix} 4 \\ 6 \\ 2 \end{pmatrix}$$

are parallel. The scalar product is

$$\begin{aligned} \vec{a} \cdot \vec{b} &= 2 \cdot 4 + 3 \cdot 6 + 1 \cdot 2 = 28 \\ &\text{and also} \\ &= \sqrt{2^2 + 3^2 + 1^2} \cdot \sqrt{4^2 + 6^2 + 2^2} \cdot \cos 0 \\ &= \sqrt{14} \cdot \sqrt{56} \cdot 1 \\ &= \sqrt{784} = 28 \end{aligned}$$

1.7.12 Vector Product

The vector product (or cross product) of two vectors

$$\vec{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \quad \text{and} \quad \vec{b} = \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix}$$

is defined as

$$\vec{c} = \vec{a} \times \vec{b} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \times \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{pmatrix}$$

The vector product only exists in three dimensions.

The resulting vector is perpendicular to both product vectors. Its direction can be found using the right hand rule, see Figure 1.9.

Example 1.17 Vector Product

$$\begin{pmatrix} 4 \\ 6 \\ 3 \end{pmatrix} \times \begin{pmatrix} 2 \\ 9 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \cdot 2 - 3 \cdot 9 \\ 3 \cdot 2 - 4 \cdot 2 \\ 4 \cdot 9 - 6 \cdot 2 \end{pmatrix} = \begin{pmatrix} -15 \\ -2 \\ 24 \end{pmatrix}$$

The length of the vector product depends on the length of the two product vectors and the angle between them:

$$|\vec{c}| = |\vec{a} \times \vec{b}| = |\vec{a}| \cdot |\vec{b}| \cdot \sin \theta$$

where θ is the enclosed angle between the two vectors. Thus,

- if the two vectors are parallel ($\theta = 0$), the length of the vector product is zero, because $\sin 0 = 0$
- if the two vectors are perpendicular ($\theta = 90^\circ$ in degrees, or $\theta = \pi/2$ in radians), the length of the vector product is just the product of the lengths of the product vectors, because $\sin 90^\circ = 1$

Example 1.18 Vector Product and Angles

$$\begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \cdot 0 - 0 \cdot 1 \\ 0 \cdot 0 - 4 \cdot 0 \\ 4 \cdot 1 - 0 \cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}$$

The length of the resulting vector is 4, and so is the product of the length of vector a (which is 4) times the length of vector b (which is 1).

The following identity is true for the triple vector product:

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

— baccab! Note that this vector is in the plane that is defined by \vec{b} and \vec{c} .

1.7.13 Matrices

A matrix looks like this:

$$\mathbf{A} = \begin{pmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{pmatrix}$$

If you multiply a matrix with a vector, you get a new vector:

$$\vec{A}\vec{b} = \begin{pmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{pmatrix} \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_{xx}b_x + a_{xy}b_y + a_{xz}b_z \\ a_{yx}b_x + a_{yy}b_y + a_{yz}b_z \\ a_{zx}b_x + a_{zy}b_z + a_{zz}b_z \end{pmatrix},$$

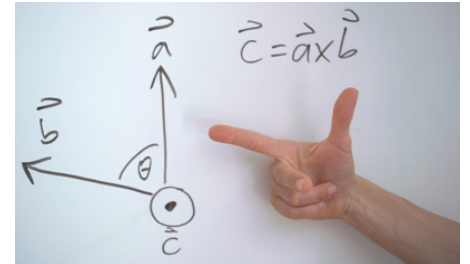


Figure 1.9: Vector Product

so each component of the new vector is the scalar product of the corresponding row in the matrix times the (column) vector.

A matrix is multiplied by a scalar by multiplying each component with that number:

$$c\mathbf{A} = c \begin{pmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{pmatrix} = \begin{pmatrix} ca_{xx} & ca_{xy} & ca_{xz} \\ ca_{yx} & ca_{yy} & ca_{yz} \\ ca_{zx} & ca_{zy} & ca_{zz} \end{pmatrix}$$

Two matrices are added by adding the components:

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{pmatrix} + \begin{pmatrix} b_{xx} & b_{xy} & b_{xz} \\ b_{yx} & b_{yy} & b_{yz} \\ b_{zx} & b_{zy} & b_{zz} \end{pmatrix} = \begin{pmatrix} a_{xx} + b_{xx} & a_{xy} + b_{xy} & a_{xz} + b_{xz} \\ a_{yx} + b_{yx} & a_{yy} + b_{yy} & a_{yz} + b_{yz} \\ a_{zx} + b_{zx} & a_{zy} + b_{zy} & a_{zz} + b_{zz} \end{pmatrix}$$

Two matrices are multiplied by doing the scalarproduct between the corresponding rows and columns:

$$\begin{aligned} \mathbf{AB} &= \begin{pmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{yx} & a_{yy} & a_{yz} \\ a_{zx} & a_{zy} & a_{zz} \end{pmatrix} \begin{pmatrix} b_{xx} & b_{xy} & b_{xz} \\ b_{yx} & b_{yy} & b_{yz} \\ b_{zx} & b_{zy} & b_{zz} \end{pmatrix} \\ &= \begin{pmatrix} a_{xx}b_{xx} + a_{xy}b_{yx} + a_{xz}b_{zx} & a_{xx}b_{xy} + a_{xy}b_{yy} + a_{xz}b_{zy} & a_{xx}b_{xz} + a_{xy}b_{yz} + a_{xz}b_{zz} \\ a_{yx}b_{xx} + a_{yy}b_{yx} + a_{yz}b_{zx} & a_{yx}b_{xy} + a_{yy}b_{yy} + a_{yz}b_{zy} & a_{yx}b_{xz} + a_{yy}b_{yz} + a_{yz}b_{zz} \\ a_{zx}b_{xx} + a_{zy}b_{yx} + a_{zz}b_{zx} & a_{zx}b_{xy} + a_{zy}b_{yy} + a_{zz}b_{zy} & a_{zx}b_{xz} + a_{zy}b_{yz} + a_{zz}b_{zz} \end{pmatrix} \end{aligned}$$

1.7.14 Nabla

This is where calculus really meets vectors: the so-called **nabla** operator

$$\vec{\nabla} = \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix}$$

This looks like a vector, and in some ways behaves like one, but of course it's not really a vector, since it neither transforms nor commutes correctly, e.g., $\vec{\nabla}\vec{a} \neq \vec{a}\vec{\nabla}$. You can have fun when you do products of this vector, e.g.,

$$\text{Gradient: } \vec{\nabla}a(x, y, z) = \begin{pmatrix} \partial_x a(x, y, z) \\ \partial_y a(x, y, z) \\ \partial_z a(x, y, z) \end{pmatrix}$$

$$\text{Divergence ("Scalar Product")}: \vec{\nabla} \cdot \vec{a}(x, y, z) = \partial_x a_x(x, y, z) + \partial_y a_y(x, y, z) + \partial_z a_z(x, y, z)$$

$$\text{Rotation ("Vector Product")}: \vec{\nabla} \times \vec{a}(x, y, z) = \begin{pmatrix} \partial_y a_z(x, y, z) - \partial_z a_y(x, y, z) \\ \partial_z a_x(x, y, z) - \partial_x a_z(x, y, z) \\ \partial_x a_y(x, y, z) - \partial_y a_x(x, y, z) \end{pmatrix}$$

This is where physics can get really interesting, for example, if one wants to understand the intertwined nature of electric and magnetic fields and the inner workings of light, etc. However, in this course, we will only get around to explore the gradient in more detail. You can really enjoy this much more once you go to physics graduate school, which of course you will immediately apply for after the end of this course.

1.7.15 Eigenfunctions and Eigenvectors

"Eigen" is German and means "one's own." Eigenfunctions and eigenvectors of operators are those which you get back with a purely numerical factor (called the eigenvalue) after applying the operator — the operator "leaves them alone."

For example, d/dx is an operator, and $A \exp(Bx)$ is an **eigenfunction** of this operator, since

$$\frac{d}{dx} A \exp(Bx) = B \cdot A \exp(Bx)$$

— you get $A \exp(Bx)$ back with an **eigenvalue** of B . Along the same lines $A \sin(Bt)$ is an eigenfunction of d^2/dt^2 with an eigenvalue of $-B^2$.

Matrices can also be considered operators. For example the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has an **eigenvector** of

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

since

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

— the eigenvalue of this particular eigenvector is -1 . As boring as these may seem, they were the key to the development of quantum mechanics.



Ask Yourself . . .

What would be the eigenvalue of the vector $\begin{pmatrix} 42 \\ -42 \end{pmatrix}$ with respect to the above matrix operator?



Ask Yourself . . .

Which other eigenvectors and eigenvalues for the above matrix operator can you find?

Problems

1.1 What is the derivative of

$$\frac{1}{2}a_0t^2 + v_0t + x_0$$

with respect to t ?

Solution: $a_0t + v_0$

1.2 What is

$$\left| \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \end{pmatrix} \right| ?$$

Solution: ≈ 53.77

1.3 Approximate $x(t) = A \sin(\Omega t)$ for small t .

Solution: $x(t) \approx A\Omega t$

1.4 With

$$\vec{F} = \begin{pmatrix} F_p \\ 0 \\ -mg \end{pmatrix} \text{ and } \vec{s} = \begin{pmatrix} l \sin \theta \\ l \cos \theta \\ 0 \end{pmatrix}$$

what is $\vec{F} \cdot \vec{s}$?

Solution: $F_p l \sin \theta$