Lecture Notes for PHY 011: Introduction to Physics

Daniel Walter Rowlands, `dwrowlands@aacc.edu`

Anne Arundel Community College, Fall 2015

Contents

1 Algebra

1.1 Order of Operations and Arithmetic with Fractions

1.1.1 Order of Operations

Since algebra is an extension of arithmetic, the same order of operations applies, although slightly extended. The algebraic order of operations is:

- First of all, any expression inside parentheses is evaluated before other operations are performed on it. For example, when evaluating $5 \times (3 + 4)$, we first evaluate $3 + 4$ to produce the expression $5 \times (7)$, which we can then evaluate as 35. This holds true even if more than one set of parentheses is used, or if parentheses are nested inside each other. In the expression $((2 + 5) \times 7) - (49 - 1)$, we first evaluate the innermost set of parentheses to get $(7 \times 7) - (49 - 1)$. We can then evaluate the remaining sets of parentheses, $49 - 48 = 1$.

- The next highest priority is exponentiation and the taking of roots. Because this has a higher priority than subtraction, if no parentheses are used, the square of a negative number is still negative: $-5^2 = -25$. However, parentheses *are* often used in this case, for example $(-5)^2 = 25$. When roots are written as non-integer exponents, they behave in the same way. But when they're written with root notation, they act as parentheses for everything under the bar: $\sqrt{5-1} = \sqrt{4} = 2$.

- After exponentiation and roots come multiplication and division. It's important to note that when an expression is given entirely above or below a fraction bar, it's treated as though it's in parentheses and evaluated before the division is. Thus $\frac{5-1}{\sqrt{4}} = \frac{4}{2} = 2$. Similarly, $-\frac{1+5}{2} = -\frac{6}{2} = -3$, but $\frac{-1+5}{2} = \frac{4}{2} = 2$.

- Finally, the lowest priority operations are addition and subtraction. If you're ever uncertain about whether to do an addition or a subtraction first, remember that you can treat subtraction as adding a negative number. This means that if you have a long sum of positive and negative numbers, such as $5 - 4 + 3 - 2 + 1$, you can replace the subtractions with additions of the negatives of the subtracted numbers: $5+(-4)+3+(-2)+1$. Then, since addition is not order-dependent, you can rearrange them however is most convenient. For example, $5 + (-4) + 3 + (-2) + 1 = 5 + 3 + 1 + (-4) + (-2) = 9 + (-6) = 3$.

1.1.2 Arithmetic Notation Used in Algebra

The notation for multiplication is slightly different in algebra than in ordinary arithmetic, even though the operations are performed in exactly the same way. The “x” multiplication sign, $\times$, is almost never used for multiplication in algebra, because it looks too much like the letters x and X, which are frequently used as variables.[1] The “dot” multiplication sign, $\cdot$, is used more frequently.

However, the most usual convention is just to omit any sign for multiplication: if two quantities are written next to each other without a sign indicating an operation, they’re assumed to be multiplied. Thus, $2 \times 4 = 2 \cdot 4 = (2)(4) = 8$. However, a physicist will almost always write $(2)(4) = 8$.

Part of the reason that physicists avoid using $\times$ or $\cdot$ to indicate multiplication of ordinary numbers is that they have different meanings when working with special quantities called vectors, as we will learn in the second half of the course. There are two different ways to multiply vectors: the “dot product”, indicated with $\cdot$, produces an ordinary number as a result; the “cross product”, indicated with $\times$ produces another vector as the result. Since vectors are very common in physics, we usually reserve the $\times$ and $\cdot$ symbols for multiplying vectors and use no symbol to indicate that we are multiplying ordinary numbers.

In addition, the division sign from arithmetic, $\div$ is almost never used in algebra and more advanced math. Instead, division is always indicated by using a fraction bar, /. For example, $5 \div 2 = 5/2 = \frac{5}{2} = 2.5$.

Writing division as fractions is useful for several reasons. One is that it simply takes up less space, in the same way that omitting a symbol for multiplication does. Another is that it reduces the number of parentheses we need to use, which can be helpful with complicated expressions. A third is that it makes canceling more intuitive: if the same term of a multiplication is on the top and bottom of a fraction, it can be canceled. That is $\frac{(2)(3)}{2} = 3$, because we can cancel out the 2s.

[1] One major exception to this comes up in scientific notation, which will be discussed in Section 2.2.

1.1.3 Working with Fractions

Since fractions come up often in algebra, it is worth quickly reviewing how to work with them. We've already discussed canceling terms present in both the numerator and the denominator. Similarly, we must remember that when multiplying two fractions, we multiply the numerators and, separately, multiply the denominators: $(\frac{2}{3})(\frac{4}{5}) = \frac{(2)(3)}{(4)(5)} = \frac{6}{20}$.

Dividing two numbers is the same as multiplying the first by one divided by the second, so $3/5 = 3(\frac{1}{5})$. This is true even if one or both of the numbers are themselves fractions. Thus, we can simplify the expression $\frac{2/3}{5/7}$ by converting it to the product of the numerator with the reciprocal of the denominator. That is, $\frac{2/3}{5/7} = (\frac{2}{3})(\frac{7}{5})$.

When adding and subtracting fractions, remember that if two fractions have the same denominator, you can combine them into one fraction with that same denominator and the sum of their numerators. Similarly, if there is a sum in a fraction's numerator, you can break it up into two fractions. That is, $\frac{3+5}{7} = \frac{3}{7} + \frac{5}{7}$. However, you *cannot* separate sums in denominators in this way.

One final note: when doing algebra, you should *never* used mixed numbers. Writing "improper fractions" like $\frac{3}{2} = 1.5$ is fine, but a mixed number like $1\frac{1}{2} = 1.5$ will be read as implying multiplication, i.e. $1\frac{1}{2} = (1)(\frac{1}{2}) = 0.5$.

1.2 The Use of Variables

1.2.1 Introduction to Variables

The most immediately obvious property of algebra is that numbers are often represented by letters, called variables.[2] Different letters (which includes upper-case and lower-case forms of the same letter, and the same letter with different subscripts) are different variables, and are assumed to be different numbers unless they are shown to be the same. For example, one cannot generally assume that $x = X$, $x = y$, or $x = x_1$. However, if $x = (3)(3)$ and $y = 4 + 5$ in a given problem, then $x = y = 9$ in that problem.

The term "variable" can be used to refer to any letter used to represent a number in algebra. However, other terms are sometimes used for certain types of variables. Variables that represent a standard value, such as the speed of light, $c \approx 3.00 \times 10^8$ m/s, or pi, $\pi \approx 3.14159$, are often called "constants." The term "constant" can also be used for any variable that doesn't change during the course of a problem, for example the weight of a cannonball being dropped from a tower. Another term used for a specific type of variable is "unknown." An "unknown" is any variable whose value is not stated at the beginning of the problem: one whose value we need to calculate.

The choice of what variables to use for what values is, in theory, up to the person writing or solving a problem. All Roman and Greek letters are considered fair game. However, letters that can easily be mistaken for numbers or other letters are best avoided. For this reason, l and I are almost never used as variables, as they can be mistaken for each other and for the number 1. When "l" is desired as a variable, it is usually written as ℓ to avoid confusion. Similarly, o and O are virtually never used, as they can easily be mistaken for the number 0.

Capital letters that look almost identical to their lower-case forms, such as P and p or X and x are also risky, although more-frequently used than I or l. Likewise, Greek letters that are identical or very similar to Roman letters, such as A (capital alpha), B (capital beta), and K (capital kappa) should generally be avoided.

Finally, it is best to avoid using certain letters as variables because they are also used to represent mathematical operations. Σ is used as the operator for summation (adding a series of numbers generated from a formula) and Π is used as the operator for the analogous operation with multiplication. Δ is frequently used to indicate change—"Δx" means "the change in x", and d is used to indicate derivatives, so it is best to avoid them as well. (However, when there is no chance of confusion, d is often used for variables representing distance.)

[2] The Grob textbook calls these "literal numbers." However, that phrase is unique to that book and is not used anywhere else in math or physics.

1.2.2 Letters Used for Specific Variables or Types of Variables

In physics, it is common to choose variables that are the first letters of the words they represent, for example m or M for a variable representing mass, or h for one representing height. In fact, physicists often use subscripts to allow all variables representing the same sort of quantity to have the same letter: in a problem with three objects, their masses will likely be m_1, m_2, and m_3. Certain letters are often associated with directions as well. By analogy to the x-, y-, and z- axes on graphs, x often represents distance in the horizontal direction and y distance in the vertical direction, or x represents east-west distance, y north-south distance, and z vertical distance.

Particularly in math, but also in physics, letters from certain parts of the alphabet are frequently used for specific purposes. Constants are frequently represented by letters near the start of the alphabet $(a, b, c, \ldots)$ while unknowns are frequently represented by letters near the end of the alphabet $(\ldots, x, y, z)$. Letters near the center of the alphabet (j, k, ℓ, m, n) are usually used to represent variables that must be integers, and are often used as subscripts. In particular, n often means "any integer". Finally, certain Greek letters $(\theta, \phi, \psi$ and their capital forms) are usually used to represent angles.

Besides these general rules, there are a few letters that virtually always represent certain specific mathematical constants and should never be used for anything else to avoid confusion. The most obvious of these is the lowercase Greek letter pi, which nearly always represents the ratio of the circumference to diameter of a circle, $\pi \approx 3.14159$. Equally important in mathematics is Euler's Constant, $e \approx 2.718$, a number that is frequently used as the base of exponential and logarithmic functions, and that is incredibly important in calculus. The third letter that nearly always represents the same important mathematical constant is the imaginary unit, $i = \sqrt{-1}$, which will be discussed further in Section 2.2.[3]

While π, e, and i should never be used as variables to represent any quantities other than the mathematical constants that they usually represent, other letters that commonly represent a specific constant in one field are often used as a general-purpose variable in another. For example, physicists usually reserve c for the speed of light in a vacuum, but it is quite common for mathematicians to use c to represent an arbitrary constant, or the third constant in a problem after a and b have been assigned. Similarly, in the physics portion of this class, g will always represent the acceleration of a falling object near Earth's surface due to gravity, $g = 9.8\ \mathrm{m/s^2}$, but in other contexts it may be used to represent many other quantities.

[3] The Grob textbook follows a convention common in electrical engineering, but not used elsewhere, of representing the imaginary unit as j to allow i to represent electrical current. However, the use of i to represent the imaginary unit is universal in math and near-universal in the sciences.

1.2.3 The Greek Alphabet

Mathematicians developed the habit of using Greek as well as Roman letters as variables in the days when ancient Greek was a basic part of the grade school curriculum, and generally taught at the same time as or before algebra. Since few students today study Greek, most students learning math and science find the Greek alphabet be just a piece of arcane trivia to memorize. However, its use is so ingrained that a familiarity with it is essential. A copy is provided here for your convenience.

Name of Letter	**Upper-Case**	**Lower-Case**
alpha	A	α
beta	B	β
gamma	Γ	γ
delta	Δ	δ
epsilon	E	ϵ
zeta	Z	ζ
eta	H	η
theta	Θ	θ
iota	I	ι
kappa	K	κ
lambda	Λ	λ
mu	M	μ
nu	N	ν
xi	Ξ	ξ
omicron	O	o
pi	Π	π
rho	P	ρ
sigma	Σ	σ
tau	T	τ
upsilon	Υ	υ
phi	Φ	ϕ
chi	X	χ
psi	Ψ	ψ
omega	Ω	ω

1.3 Polynomials and Factoring

1.3.1 Adding and Multiplying Polynomials

One of the most common sorts of algebraic expressions is the *polynomial.* A polynomial is simply a sum made up of several terms, each of which can the product of a number (called a coefficient) and one or more variables. The formal definition of a polynomial requires that the variables have positive integer exponents only: x^2 and x^5 are allowed, but $x^{1/2} = \sqrt{x}$ and $x^{-1} = 1/x$ are not. This is important when discussing functions that consist of a polynomial in x—functions of the form $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, such as $f(x) = 5x^3 + 3x^2 - 4x - 1$.

However, when it comes to simply doing arithmetic on polynomials, we can simply think of them as sums of terms without caring about the specific nature of those terms. For this purpose, we can think of a polynomial as simply having the form $(a+b+c+\ldots)$. Adding or subtracting polynomials of this sort is very simple: we simply add the individual terms.

$$(a+b) + (c+d) = (a+b+c+d)$$

or

$$(a+b) - (c+d) = (a+b-c-d)$$

Multiplying two polynomials is a little bit trickier. We need to multiply each term in the first polynomial by each term in the second polynomial; the product of the polynomials is the sum of the products of these multiplications. For the simple case of two binomials (polynomials of only two terms each), you may have learned the acronym "FOIL": First, Outer, Inner, Last as a mnemonic. The resulting polynomial is the sum of the products of the first terms in each binomial, the two outermost terms, the two innermost terms, and the last terms in each binomial:

$$(a+b)(x+y) = (ax + ay + bx + by)$$

While this can be a useful mnemonic, it doesn't extend well to longer polynomials. However, we can find the products of longer polynomials by using the distributive property of arithmetic. Since each polynomial is simply a sum, we can break the first polynomial into individual terms and multiply the second polynomial by each separately:

$$(a+b+c)(x+y+z) = a(x+y+z) + b(x+y+z) + c(x+y+z)$$

Then we evaluate each of these products to get (in this case) a nine-term polynomial:

$$ax + ay + az + bx + by + bz + cx + cy + cz$$

It's worth noting that this process works for products of more than two polynomials as well. We simply begin by selecting a pair of polynomials to multiply, then multiply the product by the third polynomial, and continue until we have gone through all of them:

$$(a+b)(c+d)(x+y) = (ac+bc+ac+ad)(x+y) = (acx+bcx+acx+adx+acy+bcy+acy+ady)$$

1.3.2 Factoring Polynomials

Factoring polynomials is the reverse of multiplying them. Instead of multiplying two or more polynomials together, we break one polynomial apart into two or more factors that, when multiplied, will yield the original polynomial as a product. Factoring is quite often useful for simplifying problems in algebra. However, while there are some specific cases where it can be done easily, there is no good universal rule for factoring polynomials of arbitrary length.

The easiest case occurs when every term in a polynomial has a common factor. For example, in $(ax + bx + cx)$, each term has x as a factor, so we can simply divide each term by x and multiply the resulting polynomial by x:

$$(ax + bx + cx) = x(a + b + c)$$

If terms contain different powers of a common factor, we can only remove the lowest power of it present:

$$(ax^2 + bx^3 + cx^4) = x^2(a + bx + cx^2)$$

Factoring a polynomial out of a larger polynomial is harder. The product of two polynomials has a term formed by the product of each possible combination of one term from each of the multiplied polynomials. However, it is possible for two or more of these terms to add to zero, disguising their original presence. This makes factoring longer polynomials at best an art and at worst impossible. For the most part, techniques for doing so are beyond the scope of this class. However, factoring polynomials that are the products of two binomials is relatively easy and worth the effort.

In the simplest case, the product of two binomials will be a four-term polynomial and you will be able to identify pairs of factors that come from each of the two polynomials. For example, in $(ax + ay + bx + by)$, we see that two terms contain a and the other two contain b, while one a and one b term contain x and the other two terms contain y. When terms have cancelled out, it is still sometimes possible to reconstruct the original binomials. One common case of this is a binomial consisting of a difference of perfect squares, $a^2 - b^2$. In this case, the presence of perfect squares tells us that both a and b were present in both binomials. That a is present as a positive square means both binomials contained a on its own; that b is present as a negative square means one contained positive b and one negative b. The other two terms have summed to zero: $(a + b)(a - b) = (a^2 - ab + ab - b^2$.

1.4 Exponentiation

1.4.1 Integer Exponents

The operation of exponentiation—raising a number to an exponent—is to multiplication as multiplication is to addition. In its simplest form, exponentiation is simply the operation of multiplying the same number times itself a number of times. Just as $a \times b$ means "add together b copies of a," a^b means "multiply together b copies of a." Since a negative number times a negative number yields a positive, if b is even, a^b will be positive for all values of a. On the other hand, if b is odd, a^b will have the same sign as a.

This simple definition makes the most sense for exponents that are positive integers: we can easily see that $5^3 = 5 \times 5 \times 5 = 125$. It is also easy enough to extend this definition to the case of 0 as an exponent: $x^0 = 1$ for any value of x because we can think of any number, x as being written $x \times 1$. If you include 0 copies of x in the product, you'll still have 1, the "multiplicative identity" (the number that can be multiplied by any other number without changing it) left over.

There is one special case, however, where this definition leads to an inconsistency, the case of 0^0. On one hand, 0 times anything—including 1—is 0, so $0^n = 0$ for all values of n. On the other hand, we earlier defined $x^0 = 1$ for all values of x. These two definitions come into conflict when we reach 0^0, so we consider 0^0 to be *undefined* in the same way division by zero is.

We can take the line of reasoning that gave us $x^0 = 1$ even further. Dividing by a number is the opposite of multiplying it so, if x^0 is 1 times no copies of x, x^{-1} is logically 1 divided by one copy of x. That is, $x^{-1} = 1/x$. More generally,

$$x^n = \begin{cases} x^{|n|} & \text{if } n \text{ is positive} \\ \frac{1}{x^{|n|}} & \text{if } n \text{ is negative} \end{cases}$$

This definition preserves the expected properties of the signs of a number raised to an even versus an odd power. Furthermore, it becomes very useful when working with fractions, particularly fractions where the numerator denominator are both polynomials, since it allows us to move factors between the numerator and denominator simply by changing the sign on the exponent.

1.4.2 Roots and Rational Exponents

The existence of the exponentiation operation implies an inverse operation that can reverse the effects of exponentiation: the taking of roots. The nth root of a number is defined as the number that, when raised to the nth power would yield that number. Thus, if $b = a^n$, then $\sqrt[n]{b} = a$.

Roots in general behave the same rules as exponents:

- Just as $a^n b^n = (ab)^n$, $(\sqrt[n]{a})(\sqrt[n]{b}) = \sqrt[n]{ab}$.
- Just as $a^n/b^n = (a/b)^n$, $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{a/b}$.
- Just as $(a^n)^m = a^{n\times m}$, $\sqrt[m]{\sqrt[n]{a}} = \sqrt[m\times n]{a}$.

However, the rule that $a^n a^m = a^{n+m}$ does not work so directly for roots. It can be retrieved however, if we define $\sqrt[n]{a} = a^{1/n}$. This extension of exponents to allow for rational numbers, rather than just integers, to be exponents is quite useful. A detailed proof of *why* is is valid is beyond the scope of this class, but it is useful to note that

$$\sqrt[n]{a^n} = (a^n)^{\frac{1}{n}} = a^{n\times\frac{1}{n}} = a^1 = a$$

as we would expect.

The identity $a^n a^m = a^{n+m}$, like the other identities for exponents does hold in this extension to rational exponents. However roots—whether expressed as traditional exponents or as rational numbers—introduce two new difficulties.

The first difficulty is the fact that most roots of rational numbers are irrational numbers; numbers that cannot be expressed as any fraction of finite real numbers. While we have already been familiar with some irrational numbers, most famously π, the four fundamental operations of arithmetic applied to the integers and rational numbers can only produce more rational numbers. Likewise, integer powers of rational numbers will always themselves be rational. However, the roots of most rational numbers are themselves irrational.[4]

The second, and more problematic, difficulty is that many roots have no real number solution. Since any real number raised to an even power is positive, it follows that the even root (square root, fourth root, etc) of a negative number cannot be a real number. The simplest solution to this is to simply treat the domain of even roots as the non-negative real numbers. However, the "complex numbers", discussed in Section 1.5 have also been developed as a means of assuring that the even roots of any real number will have a (not necessarily real) solution.

[4] In formal mathematical terminology, we say that the rational numbers are "closed" under addition, subtraction, multiplication, division, and raising to an integer power, because these operations performed on a rational number always yield a rational number. On the other hand, the rational numbers are "open" under raising to a non-integer power.

1.5 Complex Numbers

1.5.1 The Imaginary Unit

Complex numbers were originally introduced into mathematics as a fix for the problem that even roots of negative numbers cannot be evaluated as real numbers. However, the complex numbers have since been found to have many, many other useful applications in pure mathematics and they are used quite regularly in physics.

The basic postulate on which the complex number system is based is the definition of the "imaginary unit", i, which is defined as $i = \sqrt{-1}$. Factoring and the properties of exponents allow us to write any square root of a negative number in terms of i: $\sqrt{-a} = \sqrt{(-1)a} = i\sqrt{a}$, where a is a positive number. It is likewise possible to express higher even roots (fourth roots, sixth roots, and so on) of negative numbers solely in terms of real numbers and i. However, the details are too complicated to cover here.[5]

Numbers of the form bi, where b is a real number, are called "imaginary numbers". The product of an imaginary number and a real number will simply be another imaginary number: where a and b are real, $(a)(bi) = abi$. Since $i^2 = -1$, the product of two imaginary numbers will be real: $(ai)(bi) = -ab$.

Here it is useful to notice that there is a pattern in the positive integer powers of i. Since $i^2 = -1$, $i^3 = (i^2)(i) = -i$ and $i^4 = (i^2)(i^2) = (-1)(-1) = 1$. We then have that $i^5 = (1)(i) = i$, and it becomes clear that the powers of i follow the pattern $(i, -1, -i, 1, i, -1, -i, 1, \ldots)$.

This pattern in fact holds for all integer powers of i, not only the positive ones. As with anything else raised to the zeroth power, $i^0 = 1$. To find the value of $i^{-1} = 1/i$, consider that i/i should equal 1. Since $(i)(-i) = 1$, we have that $(i)(1/i) = (i)(-i)$, and thus $1/i = -i$. That $1/(i^2) = 1/(-1) = -1$ is trivial, and a similar line of reasoning to that for $1/i$ leads to the conclusion that $1/(i^3) = i$.

While it is possible to raise imaginary numbers to non-integer powers, the process for doing so is related to the matter of fourth and higher roots of negative real numbers and is likewise beyond the scope of this class.

[5] At first glance, it may seem possible to obtain a contradiction by a similar technique: $1 = \sqrt{1} = \sqrt{(-1)(-1)} = \sqrt{-1}\sqrt{-1} = (i)(i) = -1$. The error here is that, just like a positive number, -1 technically has two square roots, i and $-i$. It turns out that if we want to break $\sqrt{(-1)(-1)}$ into two separate square roots, we must use each of these roots, so $\sqrt{(-1)(-1)} = (i)(-i) = 1$. Why this is so is quite subtle, and is related to the method for finding higher even roots of negative numbers. Among other things, it turns out that any number (negative or not) has n nth roots. For more details, you can read about de Moivre's theorem.

1.5.2 The Complex Plane

While we can multiply imaginary numbers by real numbers easily enough, and can raise imaginary numbers to integer powers to produce other real or imaginary numbers, there is no way to add a real number to an imaginary. If we attempt to do so, for example by adding a to bi (where a and b are real), we are simply left with a binomial containing one real and one imaginary term. Such a number, containing a real and an imaginary term, is called a "complex number"; real and imaginary numbers are just subsets of the complex numbers that happen to have an imaginary or real part equal to 0.

The standard form for writing a complex number is $a+bi$ with a and b real numbers; any complex number can be reduced to this form. Reducing sums or products of complex numbers is simple enough: just add or multiply them as you would any other polynomials, remembering to simplify powers of i according to the pattern discussed in the previous section. Reducing fractions with complex numbers in the denominator to $a+bi$ form is somewhat more difficult, and requires making use of what are called "complex conjugates".

The complex conjugate of a complex number $a+bi$—often written $\overline{a+bi}$—is $\overline{a+bi} = a-bi$. In other words, the complex conjugate of a complex number is found by changing the sign of the imaginary part while keeping the real part the same. The reason complex conjugates are useful is that any complex number times its complex conjugate will be a real number:

$$(a+bi)(\overline{a+bi}) = (a+bi)(a-bi) = a^2 + abi - abi + (bi)^2 = a^2 - b^2$$

If we need to remove a complex number from the denominator of a fraction, we can multiply the fraction by a fraction (equal to 1) whose numerator and denominator are both the complex conjugate of the denominator of the first fraction. The product of the denominators will be a real number, and all imaginary parts will be isolated in the numerator of the product function.

Recognizing that any complex number can be written in $a+bi$ form leads to an insight about the nature of the complex numbers as a set. In the same way that the real numbers can be arranged on a line, the complex numbers can be naturally arranged in a plane, with the real part of a complex number as its x-coordinate and the imaginary part as its y-coordinate.

1.5.3 Absolute Values of Complex Numbers

Viewing the complex numbers as a plane has a number of interesting consequences, two of which are worth noting. First, since the imaginary part of a complex number is its y-coordinate, we can visualize the taking of a complex conjugate as simply reflecting the number across the x-axis. Second, this model allows us to extend the concept of the absolute value to the complex numbers.

Just as we can find the distance between two points on a normal coordinate plane using the Pythagorean theorem—a straight line between the points will form the hypotenuse of a right triangle whose sides are the differences in the points' x and y coordinates—we can use the Pythagorean theorem to find the distance between two complex numbers. If one of those complex numbers is 0 and the other is $a + bi$, then the hypotenuse has length $\sqrt{a^2 + b^2}$ In the case of a purely real number, this reduces to $\sqrt{a^2}$, which will simply supply the absolute value of a, $|a|$. As such, it makes sense to think of the distance of a complex number from the origin as simply an extension of the concept of absolute value: the absolute value of any complex number $a + bi$ is $|a + bi| = \sqrt{a^2 + b^2}$.

This is particularly useful because it gives us a meaningful way to think about how "far apart" two complex numbers are. In general, the distance between any two real numbers on the number line is the absolute value of the difference between the numbers, $|x_2 - x_1|$. That the same definition makes sense for the complex numbers, where the distance between two complex numbers z_1 and z_2 is $|z_2 - z_1|$, is evident when we consider that our definition of absolute values for complex numbers comes directly from using the Pythagorean theorem to measure the distances of complex numbers from the origin.

1.6 Logarithms

1.6.1 Logarithms and Antilogarithms

Like roots, logarithms are a class of functions developed to solve equations involving exponentiation. Starting with the equation $x = b^a$, roots are defined such that $b = \sqrt[a]{x}$. Logarithms allow us to isolate the third of the three variables in this equation: $\log_b(x) = a$. In this case, b is known as the "base" of the logarithm; we read the equation as "log-base-b of x equals a." Since this definition of the logarithm implies that $b^{\log_b(a)} = a$, the function b^x is sometimes referred to as the "antilogarithm" $\text{antilog}_b(x)$.

Logarithms have a number of uses in mathematics, some of them rather surprising, but the main reason that we will care about them—and the main reason they were originally developed[6]—is that there are a number of identities regarding them that can be useful for solving equations involving exponents and for converting between addition and multiplication:

- $\log_b(xy) = \log_b(x) + \log_b(y)$
- $\log_b(x/y) = \log_b(x) - \log_b(y)$
- $\log_b(x^p) = p\log_b(x)$
- $\log_b(\sqrt[p]{x}) = \frac{\log_b(x)}{p}$
- $b^{\log_b(x)} = x$ and $\log_b(b^x) = x$ (The definition of an antilogarithm.)
- $\log_b(b) = 1$ and $\log_b(1) = 0$
- $x^{\log_b(y)} = y^{\log_b(x)}$

It is also important to note that $\log_b(0)$ is undefined for any base b. This follows from the fact that b^x is non-zero for any value of x. Also, $\log_b(x)$ is a non-real complex number for any negative value of x; for the purposes of this class we will treat logarithms of negative numbers as nonexistent.

[6] The discussion in the Grob textbook of mantissas and characteristics is largely an obsolete relic of the days when the only way to easily calculate a logarithm was to use a log table. In particular, the treatment of logarithms of numbers less than 1 as having negative characteristics but positive mantissas is virtually never used today.

1.6.2 Common and Natural Logarithms

While it is possible to calculate logarithms with any base, two bases are used much more frequently than any others. The first tables of logarithms were drawn up for base-ten logarithms, and these logarithms are usually referred to simply as "common logs" and written $\log(x)$ rather than $\log_{10}(x)$.[7]

Common logarithms are particularly useful because our number system is base-ten. Tables of common logs were originally used as aids to calculation, by allowing one to add the logarithms of two numbers instead of multiplying the numbers themselves. (This is how slide rules work.) While calculators have made this purpose obsolete, common logarithms are still very useful for discussing data that covers a large number of order of magnitudes, and many scales of measurement, including the pH scale used to measure acidity in chemistry and the Richter scale used to measure the strength of earthquakes are what are called logarithmic scales. This means that, for example, the Richter number of an earthquake is the common logarithm of its intensity, and a Richter 5 earthquake is ten times as intense as a Richter 4 earthquake.

The other very common base for logarithms is Euler's constant, an irrational number $e = 2.718$. Logarithms with base e are called "natural logarithms" and usually written $\ln(x)$ rather than $\log_e(x)$. Natural logarithms are particularly useful because powers of e happen to have many particularly useful properties, especially in calculus. One easy-to-appreciate value of functions of the form e^x is that they can be used particularly easily for calculating the rate of growth in systems where the rate of growth is related to the size of what is growing, such as in calculating compound interest.

Most scientific and graphing calculators will have separate keys for evaluating common logs and natural logs, but no function for directly evaluating a logarithm with an arbitrary base. This is not a great hardship, because there is a simple formulate for calculating the value of a logarithm of an arbitrary base in terms of logarithms of any other base:

$$\log_b(x) = \frac{\log_k(x)}{\log_k(b)}$$

[7] To make things more confusing, while the term "common log" universally refers to base-ten logarithms, people in different fields sometimes use $\log(x)$ to refer to logarithms in different bases. In physics, chemistry, and other physical sciences, $\log(x)$ always means the common log, $\log_{10}(x)$. However, mathematicians often write $\log(x)$ for the "natural log", $\log_e(x)$. And computer scientists and programmers often write $\log(x)$ when they mean the binary log, $\log_2(x)$, because of the importance of binary numbers in computer science.

1.7 Solving Equations

1.7.1 Linear and Quadratic Equations

Perhaps the most characteristic activity in algebra is solving equations: rearranging an expression containing an equals sign and an unknown until the unknown is isolated on one side of the equals sign and its value can be directly calculated. In simple cases, this can be done almost intuitively, once one knows the rules for how one is permitted to rearrange an equation. These rules can be summarized as "Anything you do to one side of an equation, you must do to the other side as well." You can multiply both sides of an equation by a constant or variable, or divide them by a constant. (Dividing them by a variable or expression containing a variable is possible, but you must make sure that you aren't accidentally dividing by zero.) Similarly, you can add an expression to or subtract an expression from both sides of an equation. You can also raise both sides of an equation to the same power, or raise some power to both sides of the equation, or take a logarithm with the same base of both sides of the equation.

One very common sort of problem involving solving an equation is solving for the "zeros" of a function, that is the values of x such that $f(x) = 0$. If you can find the zeros of a function, you can trivially find the values of x that give it any other value by subtracting that value from the expression of the function and finding the zeros of the modified function.

Two specific sorts of functions that it is often particularly useful—and relatively easy—to find the zeros of are linear and quadratic functions. These are the two simplest cases of polynomial functions—functions that consist of polynomials where each term is a different power of x times a constant. A linear function has terms only for x^1 and x^0, while a quadratic function has terms for x^2, x^1, and x^0.

For a linear function written in the form $f(x) = mx + b$, we can find the zeros as follows. First, set $0 = mx + b$. Subtract b from each side, giving $-b = mx$. Finally, divide each side by m, yielding the solution $x = -b/m$.

The derivation of the solution for the zeros of a quadratic function is somewhat more complex, but it can also be written as a relatively simple formula. For a quadratic function of the form $f(x) = ax^2 + bx + c$, $f(x) = 0$ when

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This equation can yield zero, one, or two values for x, depending on the value of the expression $b^2 - 4ac$. If it is negative, the square root will not have a real value and there will be no real values of x produced. If it is positive, there will be positive and negative versions of the square root, yielding two values of x. And if it is zero, there will be one value for x, since $+0$ and -0 are equivalent.

1.7.2 Higher-Order Polynomial Equations

As noted in the previous section, linear and quadratic functions are the two simplest cases of polynomial functions. One can add arbitrarily higher powers of x to a polynomial function to create higher-order functions. (The "order" of a polynomial function is the highest power of x it contains; i.e. a second-order polynomial function is a quadratic function.) While formulas of the same nature as the quadratic formula can be derived for third-order (cubic) and fourth-order (quartic) equations, they are much more complicated than the quadratic formula and rarely used. Furthermore, it can be shown that no general equation of this nature can be used to find the roots of fifth-order (quintic) or higher polynomial functions.

However, in some special cases it is possible to easily find the zeros of higher-order polynomials. Two methods are worth mentioning. First, in some cases it is possible to use a substitution of variables to convert a higher-order polynomial into a quadratic polynomial in a different variable. For example, $0 = 4x^4 + x^2 + 3$ only has terms for x^0, x^2, and $x^4 = (x^2)^2$. This means that we can define a new variable, $a = x^2$ and substitute it in to get $0 = 4a^2 + a + 3$, which is quadratic in a, meaning that we can find the value of a via the quadratic formula. Once we have the up-to-two values of a from the quadratic formula, we can use the equality $a = x^2$ to find the values of x corresponding to them. Since square roots can have positive or negative values, it is possible that we will find up to four roots.

Second, factoring a polynomial can reveal some of its zeros. Since 0 times any other number equals 0, if any of the factors of a polynomial equals 0, the polynomial as a whole will equal 0. This means that if we can factor a polynomial, its zeros are simply the zeros of its various roots. For example, $f(x) = x^3 + 2x^2 + 3x$ is a cubic function, so we cannot find its zeros directly. However, if we factor out an x, we will have the product of x, which trivially has a zero at $x = 0$ and a quadratic polynomial, whose zeros we can find with the quadratic formula.

It is worth noting that in both of the above examples, the maximum number of zeros possible was equal to the order of the polynomial. This is not a coincidence: it has in fact been proven that any polynomial function can have a number of real roots as great as its order but no greater. Furthermore, if we include complex zeros (such as when the term in the square root in the quadratic formula is negative) and count repeated instances of a root repeatedly (such as when the same factor is present twice in a polynomial), a polynomial *always* has a number of zeros equal to its order.

1.8 Solving Systems of Linear Equations

1.8.1 Systems of Linear Equations

In Section 1.7, we learned to solve for a single unknown in a single equation. This technique is sufficient for finding out when a function will have a particular value: we can set the function equal to that value and then solve for x. However, it is also sometimes useful to solve for the value of x that will give two functions the same value. This amounts to determining the value of x at which the graphs of the two functions will cross.

More generally, we sometimes need to find values of several variables that can satisfy several equations simultaneously. That is, a set of values, such as $x = 5$, $y = -4$, $z = 2$, that satisfy a set of several equations containing those variables at the same time. We say a set of values satisfy an equation if we can substitute them for the variables in the equation and get a valid equation. For example, the equation $y + 1 = x^2$ is satisfied by $x = 5$, $y = 24$, but not by $x = 24$, $y = 5$.

It is particularly useful, and particularly simple, to solve systems of linear equations—equations where all variables are to the first power, i.e. x but not $\frac{1}{x}$ or x^2. As a general rule, we can solve a system of linear equations when the equations contain the same number of variables as the number of equations.

If there are more variables than equations, the system is "underconstrained": we do not have enough information to find all the variables, and more than one set of values will satisfy the equations.

If there are more equations than variables, we have more information than we need: we could solve the system while ignoring one of the equations. It is also likely that such a system of equations will be "overconstrained": that is, it will have no set of values that will satisfy it. This can also happen when we have the same number of variables as equations, but it is less likely.

A system with the same number of variables and unknowns can be put in standard form by arranging the equations with the variables on one side of the equals sign and the non-variable constants on the other side, and arranging the equations so that the variables are in the same order in each. For example, for a system of three variables and three equations:

$$a_1x + b_1y + c_1z = d_1$$

$$a_2x + b_2y + c_2z = d_2$$

$$a_3x + b_3y + c_3z = d_3$$

1.8.2 Solving by Substitution

The simplest method for solving a system of linear equations is substitution. To solve a system of equations by substitution, we first solve one of the equations for a single variable. Then, we substitute the expression we have found for that variable into another equation, reducing the number of variables present in that equation. If we are working with a system of two variable and two equations, we now have an equation containing only one variable, which we can solve for. Once this is done, we can substitute its value for the variable in either of the equations to find the value of the other equation.

For example, consider the system of equations:

$$2x + 7y = 0 \text{ and } 4x - y = 1$$

We can begin by solving the first equation for x.

$$2x + 7y = 0 \;\rightarrow\; 2x = -7y \;\rightarrow\; x = \frac{-7}{2}y$$

We can then substitute this expression for x into the second equation and solve for y.

$$4x - y = 1 \;\rightarrow\; 4(\frac{-7}{2}y) - y = 1 \;\rightarrow\; -14y - y = 1 \;\rightarrow\; y = \frac{-1}{15}$$

If we substitute $y = \frac{-1}{15}$ into the first equation and solve for x, we get:

$$2x + 7(\frac{-1}{15}) = 0 \;\rightarrow\; 2x = \frac{7}{15} \;\rightarrow\; x = \frac{7}{30}$$

As a check, we can substitute these values into both equations and confirm that they satisfy them:

$$2(\frac{7}{30}) + 7(\frac{-1}{15}) = 0$$
$$4(\frac{7}{30}) - (\frac{-1}{15}) = 1$$

In theory, this method should work even for larger systems: three equations with three unknowns, or four equations with four unknowns, or even more. In practice, it becomes more and more unwieldily to work with with larger systems. Thus, for larger systems, it is often easier to use matrices to organize the system so that solving it can be done more automatically.

1.8.3 Matrices and Determinants

Plainly speaking, matrices are just rectangular arrays of numbers. They can be of any size, measured in terms of the number of rows and columns they contain, but each row must contain the same number of columns, and each column must contain the same number of rows. For the purposes of solving systems of linear equations, however, we only need to concern ourselves with “square” matrices: matrices with the same number of rows and columns. For example,

$$\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \text{ is a 2x2 matrix and } \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \text{ is a 3x3 matrix.}$$

Every square matrix has a value called a “determinant” that can be calculated for it. This value is something akin to the absolute value of the matrix, and it is in fact indicated by using absolute value signs around the matrix.

The formula for finding the determinant of an arbitrary large matrix is complex, and the amount of work needed to calculate these determinants by hand scales up very quickly, so we will confine ourselves to 2x2 and 3x3 matrices.

The determinant of a 2x2 matrix can be found by subtracting the product of the upper-right and lower-left corners from the product of the lower-right and upper-left corners:

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - a_2b_1$$

For the determinant of a 3x3 matrix, sum the values of each of the products of the three upper-left to lower-right diagonals (wrapping around as necessary) and then subtract from them each of the products of the three lower-left to upper-right diagonals:

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1b_2c_3 + b_1c_2a_3 + c_1a_2b_3 - a_3b_2c_1 - b_3c_2a_1 - c_3a_2b_1$$

1.8.4 Using Matrices to Solve Systems of Equations

Now that we know how to evaluate the determinants of 2x2 and 3x3 matrices, we can use them to solve systems of two or three equations and unknowns. To do this, we have to define a "system matrix" for the system of equations and "variable matrices" for each of the variables. With the equations in standard form (see Section 1.8.1), we simply create a matrix consisting of the coefficients of the variables for the system matrix:

$$\begin{array}{l} a_1x + b_1y + c_1z = d_1 \\ a_2x + b_2y + c_2z = d_2 \\ a_3x + b_3y + c_3z = d_3 \end{array} \text{ yields a system matrix of } \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}$$

To find the variable matrices, we replace the columns of coefficients of each matrix with the column of constants that the equations are equal to. Thus the matrices are

$$\begin{bmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{bmatrix} \text{ for x, } \begin{bmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{bmatrix} \text{ for y, and } \begin{bmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{bmatrix} \text{ for z.}$$

We then divide the determinants of the variable matrices by the determinant of the system matrix to find the values of the variables.

$$x = \frac{|\text{x-matrix}|}{|\text{system matrix}|},\ y = \frac{|\text{y-matrix}|}{|\text{system matrix}|},\ \text{and } z = \frac{|\text{z-matrix}|}{|\text{system matrix}|}$$

The same method can be used with a 2x2 matrix as well, although it is likely not worth the effort in comparison to solving by substitution.

2 Trigonometry and Vectors

2.1 Right-Triangle Trigonometry

2.1.1 Right Triangles

Trigonometry begins with the relationships between the lengths of the sides of right triangles and their angles. As we will see, the same functions that we can define to describe the shapes of right triangles can be extended to describe any angeles and can be used for a wide variety of purposes, including calculating the products of vectors and describing periodic and circular motion.

Right triangles are particularly interesting for several reasons. First of all, since they have two perpendicular sides, the lengths of these sides correspond to the "rise" and "run" components of the slope of the hypotenuse. That is, the two non-right angles in a right triangle define the slope of the hypotenuse. Furthermore, since the internal angles of any triangle must add up to 180°, the two non-right angles must add up to 90° and a single angle is sufficient to describe the ratio of the two perpendicular sides and the slope of the hypotenuse. In other words, the entire shape of a right triangle can be described by a single angle: any two right triangles that share angles of a given measure—say 37°—will be congruent.

In addition, the Pythagorean theorem fixes the relationship between the lengths of the two perpendicular sides of a right triangle and its hypotenuse. If we label the two sides as "adjacent" to and "opposite" from an angle, as in Figure 1, the Pythagorean theorem tells us that:

$$(\text{opposite})^2 + (\text{adjacent})^2 = (\text{hypotenuse})^2$$

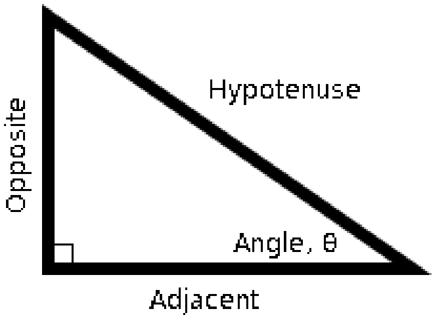

Figure 1: A right triangle, with the hypotenuse and sides opposite from and adjacent to an angle, θ, labeled.

2.1.2 The Six Basic Trig Functions

Trigonometric functions are functions that take angles as their inputs and output unitless ratios. In right-triangle trigonometry, we can define them as the ratios of the various sides of the triangle. While the hypotenuse—the side that is opposite the right angle—is uniquely defined, the other two sides, like the other two angles, are interchangeable. However, once we select which angle we are measuring, we can distinguish between them: one is opposite this angle and one is adjacent to it, as shown in Figure 1. (The hypotenuse, of course, is adjacent to both non-right angles.) The six basic trigonometric functions are then defined as ratios of the lengths of pairs of these sides as follows:

- Sine is defined $\sin(\theta) = \frac{\text{opposite}}{\text{hypotenuse}}$.
- Cosine is defined $\cos(\theta) = \frac{\text{adjacent}}{\text{hypotenuse}}$.
- Tangent is defined $\tan(\theta) = \frac{\text{opposite}}{\text{adjacent}}$.
- Cosecant is defined $\csc(\theta) = \frac{1}{\sin\theta} = \frac{\text{hypotenuse}}{\text{opposite}}$.
- Secant is defined $\sec(\theta) = \frac{1}{\cos\theta} = \frac{\text{hypotenuse}}{\text{adjacent}}$.
- Tangent is defined $\cot(\theta) = \frac{1}{\tan\theta} = \frac{\text{adjacent}}{\text{opposite}}$.

Since the second set of three functions (cosecant, secant, and cotangent) are just the reciprocals of the first set, they are rarely used, and it is common to only use the first three. In particular, calculators generally have keys for evaluating the sine, cosine, and tangent of an angle, but not the cosecant, secant, or cotangent.[8]

It is also interesting to observe that if we select the angle on the horizontal side of the triangle (its "base") to measure, then the tangent—opposite over adjacent—will measure rise over run. In other words, the tangent of the angle will describe the slope of the hypotenuse.

[8] Many calculators do have keys labeled $\sin^{-1}$, $\cos^{-1}$, and $\tan^{-1}$, but it is important to know that these *do not* calculate cosecant, secant, or cotangent. Instead, they calculate inverse trig functions, which will be described in Section 2.2.3.

2.1.3 Trig Functions of Important Angles

The most direct way to determine the trig functions of a given angle would be to use a ruler and protractor to draw a right triangle with that angle and measure the appropriate sides. This is obviously not very convenient, and the accuracy of our calculations would depend on drawing skill and the accuracy of the ruler and protractor scales. There are calculus-based ways to calculate the value of trig functions without drawing and measuring diagrams, but—besides requiring calculus—they require huge amounts of arithmetic to calculate a single value. Historically, before calculators were available to do these calculations, people generally relied on published tables of the values of trig functions for different angles.

However, there are a few particularly important angles that it is worth being able to find the trig functions of without relying on a calculator. The most obvious are 0° and 90°, which are the two limits of the angles allowed by our definition of the trig functions, since the two non-right angles in a right triangle must add up to 90°. To evaluate these cases, we have to imagine what a right triangle looks like as it approaches the extreme case of having two right angles and one angle with a measure of 0°.

If we choose the 0° angle to work with and use the adjacent side as the triangle's base, we have a hypotenuse with a slope of essentially 0: the opposite side of the triangle has a length of essentially 0 while the adjacent side and the hypotenuse, which are essentially parallel, have essentially equal lengths. This means that the sine—opposite divided by hypotenuse—is $\frac{0}{1} = 0$ while the cosine—adjacent divided by hypotenuse—is $\frac{1}{1} = 1$. Meanwhile, the tangent—opposite divided by adjacent—is $\frac{0}{1} = 0$.

Working with the 90° angle of this triangle instead gives an adjacent side of a length of essentially zero, a hypotenuse with an infinite (undefined) slope, and an opposite side essentially equal in length to the hypotenuse. This means that the sine—opposite divided by hypotenuse—is $\frac{1}{1} = 1$ while the cosine—adjacent divided by hypotenuse—is $\frac{0}{1} = 0$. Meanwhile, the tangent—opposite divided by adjacent—is $\frac{1}{0} =$ undefined.

For angles in between the extreme cases of 0° and 90°, we cannot rely on visual arguments of this sort. Instead, we can take advantage of the fact that we can evaluate the trig functions of angles of 30°, 45°, and 60°—one-third, one-half, and two-thirds of 90°—by using triangles that happen to have convenient ratios between the lengths of their sides. These triangles, shown in Figure 2 with their angles and the lengths of their sides labeled, are commonly referred to as the 45-45-90 and 30-60-90 right triangles. We can evaluate the values of trig functions at these angles by selecting the appropriate triangle and then taking the ratios of the appropriate sides.

For example, to evaluate $\sin 30^\circ$ we use the 30-60-90 triangle. Selecting the 30° angle as θ, we find that the opposite side has a length of 1 and the hypotenuse has a length of 2, so $\sin 30^\circ = \frac{1}{2}$.

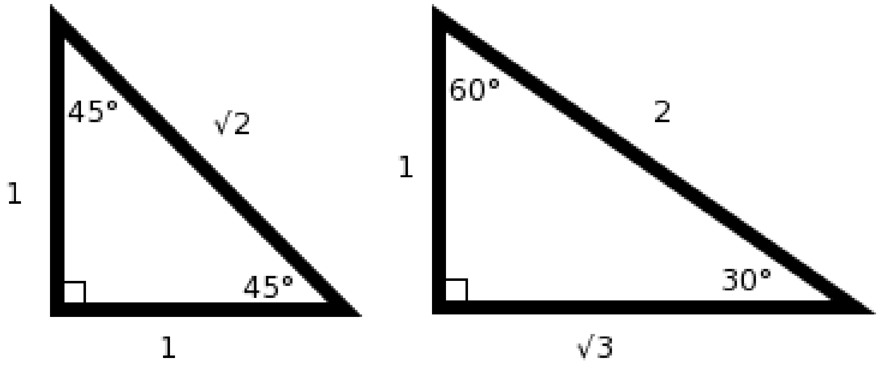

Figure 2: A 45-45-90 and a 30-60-90 right triangle, each with side lengths and angles labeled. Note that the selections of angles fixes the *ratios* of the side lengths, not the lengths themselves. Any two triangles with the same set of angles will be congruent, but, for example, one 45-45-90 triangle might have a hypotenuse of $\sqrt{2}$ and sides of length 1 while another might have a hypotenuse of length $\frac{\sqrt{2}}{\sqrt{2}} = 1$ and sides of length $\frac{1}{\sqrt{2}}$. Since the ratios of the sides remain unchanged, so do the values of trig functions of the angles.

2.1.4 Arithmetic with Trig Functions

Trig functions, like logarithms, have implied parentheses around their inputs. When taking the trig function of an algebraic expression—for example $35° + x$—we have to perform the operations described in the expression before we can evaluate the trig function. Unlike logarithms, there are not particularly simple identities for moving values into and out of the function. There are various identities that can be used to, for example, write $\sin\theta$ in terms of trig functions of 2θ, but they are very specific, and will not be used in this class. Remember that $\sin 2\theta \neq 2\sin\theta$.

One additional notational matter is important to keep in mind when working with trig functions. By common convention, positive exponents of a trig function can be—and often are—written with the exponent between the name of the function and the input of the function. That is, $(\sin\theta)^2$ is often written as $\sin^2\theta$.

This notation does *not* hold for negative exponents, however. In particular, $\sin^{-1}\theta$ refers to the “inverse sine” or “arcsine”, discussed in Section 2.2.3, and *not* to $\frac{1}{\sin\theta} = \csc\theta$.

2.2 Trigonometry on the Unit Circle

2.2.1 The Unit Circle

Defining the values of trig functions in terms of the sides of right triangles limits the domains—the sets of possible inputs—of the functions to $0° \leq \theta \leq 90°$, since angles in right triangles cannot be negative or greater than 90°. However, an angle can have a measure of any real number of degrees.

It is useful to extend our definition of the trig functions to allow us to evaluate them for any angle, not just those between 0° and 90°. Since trig functions represent the relationships between the horizontal and vertical components of the hypotenuse, we would like to be able to evaluate them for lines that slope up and down and left and right.

To do this, we imagine our angle, θ on a coordinate plane with a circle of radius 1—called a "unit circle"—and make one arm of the angle the $+x$ axis and the other arm a radius of the circle, with the radius moving counterclockwise as θ increases. Depending on the measure of angle θ, this hypotenuse will end up being in different quadrants of the coordinate plane. For $0° < \theta < 90°$, it will be in the first quadrant; for $90° < \theta < 180°$, it will be in the second quadrant; for $180° < \theta < 270°$, it will be in the third quadrant; and for $270° < \theta < 360°$, it will be in the fourth quadrant.

We can then construct a right triangle in the appropriate quadrant by using the radius as the hypotenuse, a vertical line between the radius and the x axis as the opposite side, and the section of the x axis between this line and the origin as the adjacent side, as shown in Figure 3. The angle of this triangle at the origin, ϕ, can then have its trig functions calculated.

Since the radius of a unit circle is definitionally 1, the hypotenuse will always equal 1. The opposite side will be the y coordinate of the point where the hypotenuse intersects with the circle, and so will be negative in the third and fourth quadrants, while the adjacent side will be the x coordinate of this point, and so will be negative in the second and third quadrants.

We can define trig functions of θ as being the trig functions of ϕ, taking into account that the opposite and adjacent sides will be negative when y and x are, and that this will sometimes make the trig functions of ϕ—and thus of θ—negative. Since, unlike θ, ϕ will always be between 0° and 90°, this allows us to evaluate trig functions of *any* angle in terms of the ratios of sides of right triangles. Furthermore, the values of ϕ in each quadrant are easy to determine by basic geometric rules for determining the measures of angles defined in terms of other angles.

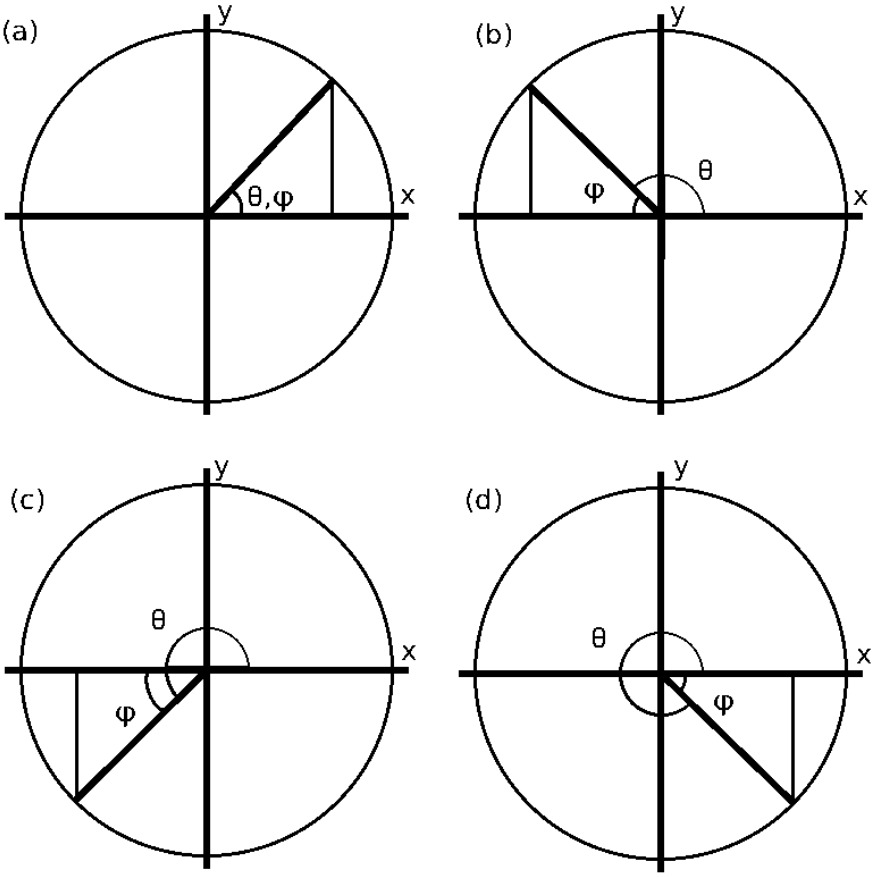

Figure 3: Angles in each of the four quadrants of the unit circle, along with right triangles using the x-axis as their bases and a radius of the unit circle as their hypotenuses.
(a) An angle θ in the first quadrant yields a triangle with an angle ϕ rising from the $+x$ axis, so the triangle's adjacent side is $+x$ and its opposite side is $+y$.
(b) An angle θ in the second quadrant yields a triangle with an angle ϕ rising from the $-x$ axis, so the triangle's adjacent side is $-x$ and its opposite side is $+y$.
(c) An angle θ in the third quadrant yields a triangle with an angle ϕ dropping from the $-x$ axis, so the triangle's adjacent side is $-x$ and its opposite side is $-y$.
(d) An angle θ in the fourth quadrant yields a triangle with an angle ϕ dropping from the $+x$ axis, so the triangle's adjacent side is $+x$ and its opposite side is $-y$.

2.2.2 Trig Functions of Arbitrary Angles

The unit circle definitions of the trig functions allow us to evaluate them for any angle between 0° and 360°. In fact, we can actually evaluate them for any angle at all, since a 361° angle will be equivalent to a 1° angle: once we loop around the circle we find ourselves back where we began. However, actually drawing out the triangles as we did in Figure 3 is somewhat time consuming, so it is useful to list out an algorithm for evaluating trig functions according to these definitions.

- If the angle is not between 0° and 360°, add or subtract an integer multiple of 360° to it to get a result that is. This is equivalent to completing a number of complete loops around the circle, each time bringing us back to where we started.
- Once we have an angle, θ, between 0° and 360°, determine which quadrant it is in and define an angle, ϕ, that is between 0° and 90°. (These rules defining ϕ can be easily derived from the method by which we defined ϕ in the previous section.)
  - If $0^\circ < \theta < 90^\circ$, we are in the first quadrant and $\phi = \theta$.
  - If $90^\circ < \theta < 180^\circ$, we are in the second quadrant and $\phi = 180^\circ - \theta$.
  - If $180^\circ < \theta < 270^\circ$, we are in the third quadrant and $\phi = \theta - 180^\circ$.
  - If $270^\circ < \theta < 360^\circ$, we are in the fourth quadrant and $\phi = 360^\circ - \theta$.
- Determine the magnitude of the trig function of θ by evaluating the same trig function for ϕ.
- Determine the sign of the trig function based on what quadrant you are in. (These rules can be memorized with the mnemonic "All Students Take Calculus", which lists which of the main trig functions—sine, tangent, and cosine—are positive in each quadrant.)
  - In the first quadrant, all trig functions are positive.
  - In the second quadrant, only sine (and thus cosecant, its reciprocal) is positive.
  - In the third quadrant, only tangent (and thus cotangent, its reciprocal) is positive.
  - In the fourth quadrant, only cosine (and thus secant, its reciprocal) is positive.

For example, let us evaluate $\sin(570^\circ)$. 570° is greater than 360°, but if we subtract 360° from it, we get $570^\circ - 360^\circ = 210^\circ$. 210° is greater than 180° and less than 270°, so it is in the third quadrant. Thus, $\phi = \theta - 180^\circ = 210^\circ - 180^\circ = 30^\circ$ and $|\sin(570^\circ)| = \sin(30^\circ)$.

Using a 30-60-90 triangle, we can see that $\sin(30^\circ) = 1/2$. Since 210° is in the third quadrant, where all trig functions except tangent (and cotangent) are negative, $\sin(570^\circ) = -1/2$.

2.2.3 Summary of Rules for Evaluating Trig Functions on the Unit Circle

Evaluating Trig Functions in Each Quadrant

Quadrant	**Expression for ϕ**	**Functions that are Positive**
Quadrant I	$\phi = \theta$	all functions
Quadrant II	$\phi = 180^\circ - \theta$	sine (and cosecant)
Quadrant III	$\phi = \theta - 180^\circ$	tangent (and cotangent)
Quadrant IV	$\phi = 360^\circ - \theta$	cosine (and secant)

Evaluating Trig Functions on Quadrant Boundaries

Function	**0°**	**90°**	**180°**	**270°**
$f(\theta) = \sin(\theta)$	0	1	0	-1
$f(\theta) = \cos(\theta)$	1	0	-1	0
$f(\theta) = \tan(\theta)$	0	undefined	0	undefined
$f(\theta) = \csc(\theta)$	undefined	1	undefined	-1
$f(\theta) = \sec(\theta)$	1	undefined	-1	undefined
$f(\theta) = \cot(\theta)$	undefined	0	undefined	0

2.2.4 Degrees and Radians

At this point, it is worth mentioning that degrees are not the only unit used to measure angles. In fact, while they are the most common unit for measuring angles in general use, they are rarely used in mathematics or physics. Instead, a unit called the radian is preferred, largely because it makes a number of things in calculus significantly simpler.[9]

Formally, the measure of an angle in radians is defined as the length of the arc in sweeps out divided by the radius of the arc. For example, an angle that sweeps out a 2-meter arc in a circle with a radius of 2 meters has a measure of 1 radian. Since the circumference of a circle is 2π times the radius of the circle, a complete rotation is 2π radians (i.e. $6.283185\ldots$ radians).

Since we also know that a complete circle is $360°$, this gives us a conversion ratio between radians and degrees: 2π radians $= 360°$, or π radians $= 180°$. Awkwardly, this means that angles that are round numbers of degrees will be irrational numbers of radians, and angles that will be rational numbers of radians will be round numbers of degrees. However, we can (and often do) leave angles in radians written in terms of π so that we can have rational values for the measures of interesting angles, such as $90°$ and $45°$. Please keep in mind, however, that π is *not* a symbol for radians, it's just $3.14159\ldots$.

To convert an angle from degrees to radians, we multiply it by the conversion ratio $\frac{\pi \text{ radians}}{180°}$. To convert from radians to degrees, we use the conversion ratio $\frac{180°}{\pi \text{ radians}}$. It is generally useful, however, to learn the measures of common angles in radians as well as degrees:

Angle in Degrees	**Angle in Radians**
$0°$	0 radians
$30°$	$\frac{\pi}{6}$ radians
$45°$	$\frac{\pi}{4}$ radians
$60°$	$\frac{\pi}{3}$ radians
$90°$	$\frac{\pi}{2}$ radians
$180°$	π radians
$270°$	$\frac{3\pi}{2}$ radians
$360°$	2π radians

[9] Mathematicians consider radians to be "unitless", because they are defined in terms of a length divided by a length. This means that several formulas—in particular for determining the rate of change of trig functions—are much simpler when angles are measured in radians than when they are measured in degrees.

2.3 Inverses of Trigonometric Functions

2.3.1 Domains, Ranges, and Periods

Now that we have defined the six trigonometric functions for arbitrary angles, we can consider the domains and ranges of these functions, and determine what their periods are. Doing so will allow us to define "inverse" functions that will reverse the effect of the trig functions. (Recall that $g(x)$ is called the "inverse of $f(x)$" if $g(f(x)) = x$.)

The "domain" of a function is defined as the set of all values that are valid inputs for a function. For example, $f(x) = x^2$ has a domain of all real numbers, since we can square any real number. On the other hand, the range of $f(x) = \sqrt(x)$ has a domain of non-negative real numbers, since we can evaluate the square root of any positive number or of zero, but not of a negative number (at least if we limit ourselves to the real numbers). Similarly, the domain of $f(x) = 1/x$ is all real numbers except zero, since division by zero is undefined.

The "range" of a function is defined as the set of all possible numbers that can be produced as outputs of a function. For example, $f(x) = x$ has a range of all real numbers, since we can make $f(x)$ equal to any real number simply by setting x equal to that number. However, $f(x) = x^2$ has a range of non-negative real numbers, since we can make $f(x)$ equal to any positive number or by setting x equal to the square root of that number, but we cannot make it equal to a negative number at all. (Consider that the graph of $f(x) = x^2$ is a parabola that opens upward and never crosses below the x-axis.)

A function is called "periodic" if there is some number n such that $f(x) = f(x + n)$ for all values of x. In other words, if there is some amount that if you increase x by, the function will repeat itself exactly. The types of functions we have discussed previously—polynomials, roots, exponentials, and logarithms—are not periodic. They may take on the same value of $f(x)$ for multiple values of x—for example, $x^2 = 4$ for both $x = -2$ and $x = 2$, but they don't take on the same value of $f(x)$ infinitely many times for each value of x.

Trig functions are the first periodic functions that we have discussed. Since traveling 360° around a circle brings you back where you started, trig functions on the unit circle all repeat themselves at least once every 360°. We call the amount that you have to increase x by to get the same value of $f(x)$ a second time the "period" of $f(x)$. While we can see that all six trig functions must repeat with a period of 360°, it is possible that they have a shorter period, if there is some smaller number that we can increase x by to get the same value of the trig function.

2.3.2 The Domain, Range, and Period of Sine and Cosine

Let us begin by determining the domain, range, and period of the sine and cosine functions. These functions are similar in that, as defined on the unit circle, where the hypotenuse is 1, they are simply expressible as $\sin(\theta) = y$ and $\cos(\theta) = x$. As we go around the unit circle from 0° to 360°, these expressions will be valid at each value of θ, so the domains of the functions are simply all real numbers.

As for the ranges of the functions, we know that as we go from 0° to 90°, x goes from 1 to 0 and y goes from 0 to 1. Since the values of trig functions in the second, third, and fourth quadrants are simply the values of the same trig function in the first quadrant times 1 or -1, we can conclude that the range of sine and cosine is $-1 \leq \theta \leq 1$. That is, -1 is less than or equal to θ and θ is less than or equal to 1: θ can have any value between -1 and 1.

As for the period of sine and cosine, as we saw in Section 2.3.1, all trig functions repeat themselves with a period of 360°. Since sine is positive in the first and second quadrants and cosine is positive in the fourth and first quadrants, both are positive for 180° and then negative for 180°. This means that they cannot have periods less than 360°. If $\sin(\theta) = \sin(\theta + n)$ for any n less than 360°, there would have to be some values of θ where $\sin(\theta)$ was positive and $\sin(\theta + n)$ was negative, which is nonsensical.

2.3.3 The Domain, Range, and Period Tangent and Cotangent

For tangent and cotangent, we can again begin with the unit circle, where $\tan(\theta) = y/x$ and $\cot(\theta) = x/y$. Since $x = 0$ when θ is 90° or 270° and $y = 0$ when θ is 0° or 180°, $\tan(\theta)$ will be undefined due to division by zero when θ is 90° or 270° and $\cot(\theta)$ will be undefined due to division by zero when θ is 0° or 180°.

Other than the cases of division by zero, y/x and x/y should be valid for all values of x and y, and thus for all values of θ. So the domain of $\tan(\theta)$ is all real numbers except $\theta = 90° + n180°$ where n is an integer, and the domain of $\cot(\theta)$ is all real numbers except $\theta = 0° + n180°$ where n is an integer.

As for the range of tangent and cotangent, we know that $\tan(0°) = 0$, $\tan(90°) = \infty$, and that tangent increases continuously between these two values. Since in the second quadrant, tangent is evaluated as $-\tan(\phi)$ where ϕ varies from 0° to 90°, we can see that the range of tangent is all real numbers. Since cotangent is the reciprocal of tangent, it too has a range of all real numbers.

Since tangent and cotangent are positive in the first and third quadrants, and since ϕ increases from 0° to 90° in those quadrants while it decreases from 90° to 0° in the quadrants where tangent and cotangent are negative, the functions have periods of just 180°.

2.3.4 The Domain, Range, and Period of Cosecant and Secant

Since $\csc(\theta) = 1/\sin(\theta)$ and $\sec(\theta) = 1/\cos(\theta)$, the properties of sine and cosine are a good starting point for determining the properties of cosecant and secant. We can conclude that the periods of cosecant and secant are 360° since the reciprocal of a function should repeat when and only when the function itself repeats.

Recall that the domains of sine and cosine are all real numbers and their ranges are from -1 to 1. The only number in this range that cannot be in the denominator of a fraction is 0, so the domains of cosecant and secant will be all real numbers except for those where sine (for cosecant) or cosine (for secant) would yield a value of 0. Since sine is positive in the first and second quadrants, its sign changes from negative to positive at 0° and from positive to negative at 180° and it equals zero at these values. Since cosine is positive in the fourth and first quadrants, its sign changes from negative to positive at 270° and from positive to negative at 90° and it equals zero at these values.

However, it is not sufficient to simply specify the two values at which cosecant and secant are undefined between 0° and 360° since the functions are periodic and the undefined points repeat themselves once every period. Instead, we have to specify an expression for all of the values of θ where the functions are undefined. We do this as follows: the domain of cosecant is all real numbers except $\theta = 0^\circ + n180^\circ$ where n is an integer; the domain of secant is all real numbers except $\theta = 90^\circ + n180^\circ$ where n is an integer.

As for the ranges of cosecant and secant, we can derive these from the properties of sine and cosine as well. Both sine and cosine have ranges of -1 to 1. The reciprocal of any number with a magnitude greater than 1 will be a number with a magnitude less than 1, so it follows that any number with a magnitude greater than 1 can be produced by taking the reciprocal of a number between -1 and 1. Thus, the range of cosecant and secant is all real numbers *except* numbers between -1 and 1.

2.3.5 Inverse Trig Functions on Right Triangles

Now that we understand the domains and ranges of the six trigonometric functions, we can think about defining their inverses. Recall that the definition of the inverse, $g(x)$, of a function $f(x)$ is that $g(f(x)) = x$. This means that an inverse function converts a number from the output range of the original function into a number in the input domain of the original function.

It is easiest to understand the geometric meanings of inverse trig functions if we first only consider the inverses of trig functions as defined on right triangles. In this case, the domain of each trig function is the set of angles from 0° to 90°, so the range of outputs for each function will be this range of angles. Since an ordinary trig function takes an angle in this domain and yields a ratio of the lengths of the sides of a right triangle containing that angle, an inverse trig function will take as an input a ratio of the lengths of the sides of a right triangle and yield an angle in that range as a result.

The six inverse trig functions are generally named by adding the prefix "arc" to the names of the trig functions they are inverses of: arcsin, arccos, arctan, arccsc, arcsec, and arccot.[10] To evaluate an inverse trig function, we have to determine what right triangle and angle would produce the ratio of sides indicated. For example, $\arcsin(\frac{2}{\sqrt{3}})$ means that we are looking for an angle where the opposite side has length 2 and the hypotenuse has length $\sqrt{3}$. We can recognize this as a 30-60-90 triangle and a 60° angle, so $\arcsin(\frac{2}{\sqrt{3}}) = 60°$. Likewise, arccot(1) means that we are looking for an angle where the opposite and adjacent sides are equal (and so have a 1-to-1 ratio). This gives us a 45-45-90 triangle and 45° angle, telling us that $\text{arccot}(1) = 45°$.

It is important to remember that the input of an inverse trig function is always a unitless ratio of the sides of a triangle, and that its output is always an angle measured in degrees (or other angle units, such as radians). Furthermore, the input of an inverse trig function is limited by the range of possible results from that trig functions. For example, arcsin(2) and $\text{arccsc}(\frac{1}{2})$ are undefined, since sine never has an output greater than 1 and cosecant never has an output in the range -1 to 1.

[10] Unfortunately, this is not the only notation used for them. They are also often written $\sin^{-1}$, $\cos^{-1}$, and so on. We will not use this alternate nomenclature in this class, because it conflicts with the common nomenclature of writing the square of a trig function as $\sin^2$, $\cos^2$, and so on. It is important to remember that inverse trig functions have nothing to do with reciprocals.

2.3.6 Full Domains of Inverse Trig Functions

Defining trig functions solely in terms of right triangles, as we did in the previous section, leaves them undefined for negative values, even though each trig function has as many valid negative outputs as valid positive outputs.

Extending the domain of the inverse trig functions to the negative numbers by using the unit circle raises a problem, however. When defined beyond the first quadrant, the trig functions are not monotonic: there are multiple values of θ that will yield the same $f(\theta)$. In order to avoid violating the vertical line test, we need to limit the output ranges of the inverse trig functions to a region in which the trig functions they are inverses of only yield each value once.

Since each trig function goes through its full range of positive values exactly once in a single quadrant, and its whole range of negative values exactly once in a single quadrant, we can simply select two quadrants to be the output range of each inverse trig function. Then, to evaluate an inverse trig function, we can use the sign to determine which quadrant to use and then determine the value of ϕ—the angle between the x-axis and the hypotenuse of a triangle in that quadrant. Then we find θ by working backwards from the formulas we used to find ϕ from θ for each trig function in that quadrant.

Since cosine and cotangent are zero at 90°, negative in the second quadrant, and continuous from 0° to 180°, it makes the most sense to define the output ranges of arccosine and arccotangent as 0° to 180°: positive in the first quadrant and negative in the second.

Sine and tangent, on the other hand, are zero at 0°, negative in the fourth quadrant, and continuous from $-90°$ to 90°, so it makes more sense to define the output ranges of arcsine and tangent as $-90°$ to 90°: positive in the first quadrant and negative in the fourth quadrant.

Secant and cosecant are more complicated, since their negative and positive regions aren't continuous. Arcsecant and arccosecant are often only defined in the first quadrant as a result. However, we can define arcsecant from 0° to 180°, positive in the first quadrant and negative in the second—just like cosine, its reciprocal—and arccosecant from $-90°$ to 90°, positive in the first quadrant and negative in the fourth—just like sine, its reciprocal. Note that in this case, 0° (for arccosecant) or 90° (for arcsecant) is a hole in the function's range, since the cosecant and secant have holes in their domains at these places.

2.3.7 Summary of the Properties of Trig and Inverse Trig Functions

Properties of Trig Functions

Trig Function	Domain	Range	Period
$f(\theta) = \sin(\theta)$	all real numbers	$-1 \leq f(\theta) \leq 1$	360°
$f(\theta) = \cos(\theta)$	all real numbers	$-1 \leq f(\theta) \leq 1$	360°
$f(\theta) = \tan(\theta)$	all real numbers $\neq 90° + n180°$	all real numbers	180°
$f(\theta) = \csc(\theta)$	all real numbers $\neq 0° + n180°$	$f(\theta) \leq -1$ and $1 \leq f(\theta)$	360°
$f(\theta) = \sec(\theta)$	all real numbers $\neq 90° + n180°$	$f(\theta) \leq -1$ and $1 \leq f(\theta)$	360°
$f(\theta) = \cot(\theta)$	all real numbers $\neq 0° + n180°$	all real numbers	180°

Properties of Inverse Trig Functions

Inverse Trig Function	Domain	Range
$\theta = \arcsin(f(\theta))$	$-1 \leq f(\theta) \leq 1$	$-90° \leq \theta \leq 90°$
$\theta = \arccos(f(\theta))$	$-1 \leq f(\theta) \leq 1$	$0° \leq \theta \leq 180°$
$\theta = \arctan(f(\theta))$	all real numbers	$-90° < \theta < 90°$
$\theta = \text{arccsc}(f(\theta))$	$f(\theta) \leq -1$ and $1 \leq f(\theta)$	$-90° \leq \theta < 0°$ and $0° < \theta \leq 90°$
$\theta = \text{arcsec}(f(\theta))$	$f(\theta) \leq -1$ and $1 \leq f(\theta)$	$0° \leq \theta < 90°$ and $90° < \theta \leq 180°$
$\theta = \text{arccot}(f(\theta))$	all real numbers	$0° < \theta < 180°$

2.4 Vectors

2.4.1 Vectors and Scalars

In physics, we generally work with two different sorts of numbers, “vectors” and “scalars”. Scalars are ordinary numbers of the type we are used to: 3, -1, π, and so on. They simply measure a quantity of something. Vectors, on the other hand, are used to measure a quantity that has an associated direction. One can think of them as representing an arrow that has some fixed length and is pointing in some fixed direction.

In general, some quantities are more naturally measured in scalars while others make more sense as vectors. Mass, for example, is a scalar quantity: it doesn’t make sense to say that a weight has a mass of two pounds in a given direction.

On the other hand, speed makes more sense as a vector. While we can specify a speed without a direction—as a car’s speedometer does—it is more useful to indicate what speed we are going and what direction. For example, we might say that we are going fifty miles per hour north, which will get us somewhere rather different than if we drove at fifty miles per hour west. In fact, vectors in different directions can cancel out. If we walk ten miles north and then ten miles south, we end up exactly where we started, the same as if we’d walked no distance at all.

While many of the same mathematical operations can be performed on vectors as on scalars, there are some differences, and it is important to know whether a variable represents a vector or a scalar. To distinguish between the two, a right-pointing arrow is generally placed above a variable that represents a vector. Thus, x is a scalar but $\vec{x}$ is a vector.[11]

A given vector is said to have a “dimension” based on how many dimensions it points in. For example, a vector that describes motion on a flat surface–where you can go north-south or east-west—would be two-dimensional, while one that describes motion of a fish—which can swim up-and-down as well as north-and-south and east-and-west—would be three-dimensional.

For the purposes of this class, we will mostly only care about two- and three- dimensional vectors, but it is important to recognize that other numbers are possible: a one-dimensional vector simply uses its sign to indicate direction—for example, motion of a train along tracks—while a four-dimensional vector is useful in mathematically-constructed spaces with more than three dimensions.

[11] This is not the only convention possible. Some texts prefer to write vectors in bold instead of using an arrow, because this is easier to type on many computers. By that convention, x is a scalar but $\mathbf{x}$ is a vector.

2.4.2 Vectors in Magnitude-Angle Form

In order to do math with vectors, we need to be able to express their values: while we could draw an arrow indicating the direction and label it with the length of the vector (the distance or speed or whatever it represents), this is not very practical.

There are several ways in which we can express the value of a vector, which we will call "forms" of the vector. It is important to remember that while a vector looks different in different forms, they are mathematically equivalent and a vector can be mathematically manipulated identically in any form. This is equivalent to the fact that your height is the same physical quantity whether you measure it in inches or centimeters: if you perform the same mathematical operation on it in either units, the effect will be the same.

The most intuitive way to express a vector is by stating its length—commonly called its "magnitude"—along with an angle or angles indicating the direction it is pointing. The magnitude of the vector is written as the name of the vector surrounded by double absolute value signs: $\|\vec{x}\|$ means "the magnitude of $\vec{x}$".

The number of angles needed to specify the direction of a vector depends on the number of dimensions the vector has. A two-dimensional vector can be specified with just one angle, θ, the same way a point is indicated in polar coordinates. Just as with trig functions, we define $\theta = 0°$ when the vector is pointing in the $+x$ direction and increase θ as it rotates counterclockwise around the origin, so that an angle with $\theta = 180°$ is pointing in the opposite direction. For example, the vector $(5,\ 90°)$ has a magnitude of 5 and is pointing in the $+y$ direction.

It is also possible do write three-dimensional vectors in magnitude-angle form. In this case we need two angles, as in spherical coordinates. Generally, one is defined in the x-y plane—like θ for two-dimensional vectors—and one is defined as the angle the vector is pointing above or below that plane. However, the mathematics becomes somewhat more complicated, and most of the physical systems we will be discussing in this class will be two-dimensional, so we will not discuss magnitude-angle form for three-dimensional vectors further.

Magnitude-angle form is a very intuitive way to think about vectors: at a glance, it allows us to read off both their magnitudes and what direction they are pointing in. However, this form makes some mathematical operations on vectors more difficult, so it is often more useful to use component form, discussed in the next section.

2.4.3 Vectors in Component Form

Another way to describe vectors is suggested by Cartesian (x-y) coordinates. If we set the start of a vector at the origin, we can describe both its length and direction by noting the coordinates of its tip. These coordinates are called "components"; they represent the portion of the magnitude of the vector that is pointing along each of the axes. A vector pointing along the line $y = x$ could have components (1, 1) or (7, 7) or so on, depending on the magnitude of the vector.[12]

For two-dimensional vectors in component form, it is fairly clear how we can find their magnitudes: simply use the Pythagorean Theorem. Since the x and y components are perpendicular, we can make a right triangle using them as sides and the vector as the hypotenuse, so

$$\|(x, y)\| = \sqrt{x^2 + y^2}$$

Interestingly—and usefully—this works for any number of dimensions. The magnitude of any vector is the square root of the sums of the squares of all its components, so

$$\|(x, y, z)\| = \sqrt{x^2 + y^2 + z^2}$$

and

$$\|(x, y, z, w)\| = \sqrt{x^2 + y^2 + z^2 + w^2}$$

and so on.

At this point, it is also worth mentioning an alternate format for writing vectors in component form. Instead of listing the components in order in parentheses—(x-component, y-component, z-component)—they can also be given as the sums of the components times "unit vectors." Unit vectors are vectors of length 1 that point along the axes in the positive directions.

By convention, unit vectors are written with a caret or "hat" above the letter instead of an arrow, and are represented by the variables $\hat{\imath}$ for the x unit vector, $\hat{\jmath}$ for the y unit vector, and $\hat{k}$ for the z unit vector. This means that the vector (x, y, z) can also be written as $x\hat{\imath} + y\hat{\jmath} + z\hat{k}$.

[12] This discussion of placing a vector on the coordinate grid leads to one possible source of confusion. It is important to understand that vectors do not have a location: they have a length and a direction, but you can set the starting point wherever you want without changing the vector. Components could be extracted from a vector by placing it anywhere on the Cartesian grid if we measured the difference in the x and y positions of the tip and base.

2.4.4 Converting between Component and Magnitude-Angle Forms

Calculating the magnitude of a vector in component form, as discussed in the previous section, brings us halfway to converting such a vector to magnitude-angle form. To find the angle as well, we'll need to use trigonometry. Since we are only using magnitude-angle form for two-dimensional vectors in this class, we can limit ourselves to finding the single angle, θ, that defines their direction.

Recall the right triangle we used to determine the magnitude of a two-dimensional vector, with the x and y components as sides and the vector as the hypotenuse. Since one corner—the base of the vector—is at the origin, and since we define angles from the $+x$ axis, we find that for a vector in the first quadrant, θ will be the interior angle of the triangle at the origin. The "opposite" side will be the y component and the "adjacent" side will be the x component, so $tan(\theta)$ will equal y/x. We can then use the inverse tangent function to find theta:

$$\theta = \arctan(y/x) \text{ in the first quadrant}$$

If either the x or y components is negative, however, this becomes

$$\phi = \arctan(|y|/|x|)$$

where ϕ is the angle between the x axis and the vector, the same as the ϕ that we used when discussing unit circle trigonometry in Section 2.2. We can then find θ as we did there, by using arithmetic to convert ϕ to an angle from the $+x$ axis. How we do so depends on what quadrant we're in:

- In the first quadrant, $\theta = \phi$, so $\theta = \arctan(|y|/|x|)$.
- In the second quadrant, $\theta = 180° - \phi = 180° - \arctan(|y|/|x|)$.
- In the third quadrant, $\theta = \phi + 180° = \arctan(|y|/|x|) + 180°$.
- In the fourth quadrant, $\theta = 360° - \phi = 360° - \arctan(|y|/|x|)$.

Unsurprisingly, converting a vector from magnitude-angle form back to component form also requires trigonometry. Using the same right triangles we used to convert from component form to magnitude-angle form, we can see that the y component is the opposite side and the x component is the adjacent side, so, for an angle (r, θ),

$$\cos(\theta) = \frac{x}{r} \to r\cos(\theta) = x$$

$$\sin(\theta) = \frac{y}{r} \to r\sin(\theta) = y$$

Since we have already defined sine and cosine in all four quadrants, these formulas will work regardless of the value of θ.

2.4.5 Arithmetic on Vectors

Now that we know how to write vectors, it is time to learn how to perform arithmetic operations with two vectors or a vector and a scalar.

The only arithmetic operation allowed between a vector and a scalar is multiplication. One cannot add or subtract a vector and a scalar, and one cannot write a vector in the denominator of a fraction, so one cannot divide by a vector. (Dividing a vector by a scalar is allowed, because it is equivalent to simply multiplying the vector by the reciprocal of the scalar.)

Multiplying a vector by a scalar simply has the effect of multiplying the magnitude of the vector by that scalar without changing the direction of the vector. In other words $\|a\vec{B}\| = a\|\vec{B}\|$. If the vector is written in magnitude-vector form, this is clearly equivalent to saying that $a(r, \theta) = (ar, \theta)$.

In component form, we get the same effect by multiplying the scalar by each component separately: $a(x, y, z) = (ax, ay, az)$. This will not change the direction of the vector, since $\frac{ay}{ax} = \frac{y}{x}$ and it will have the same effect as multiplying the magnitude by r, since

$$\sqrt{(ax)^2 + (ay)^2 + (az)^2} = \sqrt{a^2(x^2 + y^2 + z^2} = a\sqrt{x^2 + y^2 + z^2}$$

As for arithmetic involving two vectors, here both addition and subtraction are allowed, so long as the vectors are of the same dimension (two two-dimensional vectors or two three-dimensional vectors, but not one two-dimensional vector and one three-dimensional vector).

Geometrically speaking, adding two vectors means lining them up one after another and making a new vector that connects the base of the first vector to the tip of the second vector. In other words, finding one vector that combines the effects of the two vectors.

In component form, we can achieve this by adding the components along each axis separately:

$$(x_1, y_1, z_1) + (x_2, y_2, z_2) = (x_1x_2, y_1y_2, z_1z_2)$$

If we need to subtract one vector from another, we can just make use of the fact that subtraction is addition of a negative and treat the subtraction as multiplication by -1 followed by addition of vectors.

Adding two vectors in magnitude-angle form is more complicated, and we won't discuss it in this class. Instead, if you have to add two vectors in magnitude-angle form, convert them to component form and add them that way. In particular, note that adding two vectors will *not* simply add their magnitudes, unless the two angles have identical values of θ.

As with the case of dividing a scalar by a vector, we cannot divide a vector by a vector, since in general vectors don't have reciprocals and we can't put a vector in the denominator of a fraction.

2.4.6 Dot Products

The multiplication of vectors is somewhat complicated, as several different operations for multiplying two vectors—"products"—are defined. The most common products used in physics are the "dot product" (also called the "scalar product" or "inner product"), which is written $\vec{A} \cdot \vec{B}$, and the "cross product" (also called the "vector product"), which is written $\vec{A} \times \vec{B}$.

As their alternate names suggest, the dot product of two vectors is always a scalar, while the cross product of two vectors is always a vector. The dot product is defined for any pair of vectors of the same dimension (so two two-dimensional vectors or two three-dimensional vectors, but not a two-dimensional and a three-dimensional vector) while the cross product is *only* defined for a pair of three-dimensional vectors.

For vectors in component form, the dot product is calculated by multiplying the values of each component separately and summing the results:

$$(x_1, y_1, z_1) \cdot (x_2, y_2, z_2) = x_1x_2 + y_1y_2 + z_1z_2$$

For for vectors in magnitude-angle form, this is equivalent to multiplying their magnitudes by the angle between them:

$$(r_1, \theta_1) \cdot (r_2, \theta_2) = r_1r_2 \cos(\theta_2 - \theta_1)$$

Since cosine is an even function, the sign of $\theta_2 - \theta_1$ doesn't matter, and thus it doesn't matter which vector is written first.

From these definitions, we can see that the dot product of two vectors with the same direction will just be the product of their magnitudes, while the dot-product of two perpendicular vectors will be zero.

We can also see that a vector dotted with itself, often written $\vec{A} \cdot \vec{A} = \vec{A}^2$ will just be the magnitude of the vector squared: $\vec{A}^2 = \|\vec{A}\|^2$ However, exponentiation in general is not well-defined for vectors, and it is better to avoid writing a vector to any power other than 1 or 2.

2.4.7 Cross Products

As noted in the previous section, the dot product of two vectors is zero if the vectors are perpendicular and is the product of their magnitudes if they have the same direction. The exact opposite is true with cross-products, however. The magnitude of the cross product of two vectors is $\|\vec{A} \times \vec{B}\| = \|\vec{A}\|\|\vec{B}\| \sin(\theta)$, where θ is the angle between the two vectors, which is zero for vectors with the same direction and equal to the product of magnitudes for perpendicular directions.

As for the direction of a cross-product—remember that, unlike the dot product, the cross product yields another vector—it will always be perpendicular to the two vectors crossed, following something called the "right hand rule". This rule says that if you point the fingers of your open hand along the first vector, curl them in the direction of the second vector to make a fist, and then stick out your thumb in a "thumbs-up" sign, the product vector will be aligned along the direction of your thumb.

Because the direction of the product vector is defined this way, cross products—unlike dot products—are order-dependent. Switching the order of the two vectors will switch the direction of the product vector by 180°, so $\vec{A} \times \vec{B} = -\vec{B} \times \vec{A}$.

Since cross products are only defined in three dimensions, we will only discuss evaluating them in component form. The formula for evaluating cross products is often written in the form of the determinant of a matrix:

$$(x_1, y_1, z_1) \times (x_2, y_2, z_2) = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}$$

This determinant is equivalent to

$$(x_1, y_1, z_1) \times (x_2, y_2, z_2) = ((y_1 z_2 - z_1 y_2),\ (z_1 x_2 - x_1 z_2),\ (x_1 y_2 - y_1 x_2))$$

3 Physical Quantities and Units of Measurement

3.1 Uncertainty and Significant Figures

3.1.1 Uncertainty

Physics has—not unreasonably—been described as just being math where you give particular names to the variables and associate them with things in the real world. One consequence of this association is that numbers in physics problems almost always represent measured quantities: if we say something weighs twenty pounds, or is five feet long, or that the air temperature is ninety degrees Fahrenheit, we mean that someone took out an instrument—a scale or a measuring tape or a thermometer—and measured it.

This makes numbers in physics fundamentally different from numbers in math because numbers that come from measurements always have some uncertainty in them. If I tell you that I want to know the square root of four, or the solution to $y = 7x - 3.45$, you can assume that I mean exactly four, exactly seven, and exactly -3.46.

However, if I tell you that an object weighs twenty pounds, is five feet long, and has a temperature of ninety degrees Fahrenheit, you know those numbers aren't exact. Maybe the scale, measuring tape, and thermometer were miscalibrated. Maybe I took my measurements sloppily: put my finger on the scale or didn't hold the tape right to the end of the object. But even if I did everything perfectly, the readings these instruments are limited by their precisions. If the measuring tape is marked every quarter-inch, I don't really know if the object is five feet and a sixteenth of an inch, or five feet even, or five feet and an eighth of an inch. If the scale is only marked in pounds and ounces, I don't really know if my reading is off by a quarter ounce.

In order to take into account this uncertainty in the values of numbers that come from measurements, physicists often associate numbers with an "uncertainty" or "error bar", usually written in the form of a plus-or-minus sign and then a second number after the measurement result, indicating a certain certainty—often 95%—that the true value of the thing they're measuring is between the number given minus the uncertainty and the number given plus the uncertainty.

This technique allows a lot of freedom to specify how uncertain the physicist is about a result, and whether it is likely that a second, similar, measurement might actually be identical (if they are within each others' uncertainty). However, quantifying the uncertainty in a measurement is a difficult process that depends on a knowledge of significant statistics and the details of the instrument one is using to take the measurement.

Furthermore, figuring out the uncertainty of the output of a mathematical expression—what should the uncertainty of $\sqrt{5.0 \pm 0.1}$ be?—is itself rather complicated and depends on a knowledge of calculus, so we will not use uncertainties specified in this way.

3.1.2 Significant Figures

Instead of dealing with the detailed statistics of precisely quantifying uncertainty, we will keep track of uncertainty using a set of general rule-of-thumb rules to determine how many digits to round our results to in order to maintain roughly the same level of uncertainty as was present in the numbers we used to find them. These rules will ensure that our answers indicate a reasonable degree of uncertainty, neither throwing away meaningful information nor containing long decimals that mean little because our uncertainty keeps us from knowing the value of the digits.

The rules we will use, which are common in science in general, are called "significant figures". The basic principle is that the uncertainty of a number can quantified by how many reliably-known digits it contains. These digits, along with the first digit in which there is some uncertainty, are called "significant figures", and our goal is to write our answer with as many significant figures as are available to us, given the information provided, but no more.

Essentially, all non-zero digits in a number are significant, as are zeros between two non-zero digits or a non-zero digit and a decimal point. Thus:

- 2406.01 has six significant figures.
- 2406.0 has five significant figures.
- 2400 has two significant figures
- 2400. has four significant figures
- 0.01 has two significant figures.

This only applies to numbers derived from experimental measurements with instruments, however. Integers derived from exact counting—if you say you have three sisters, you mean exactly three, not 3.0 ± 0.5—as well as conversion ratios between units that are defined in terms of each other—one foot is exactly twelve inches—and mathematical constants like π are said to have infinite significant figures, and they don't constrain the certainty of your answers.

Once you know how many significant figures are available in the numbers you are working with, you can use the rules in the next section to keep track of how many significant figures will be present in your answer. When you get to your final answer—but not before—you should round it to the correct number of significant figures.

Remember: Don't round off to sigfigs until your final answer. Otherwise, you risk introducing rounding errors.

3.1.3 Rules for Significant Figures

- Sigfigs are a *roundoff* rule. That means that you use them for rounding your final answer. However, you don't want to round intermediate values because this adds rounding error to your work. *Only round the final result.*

- You need to pay attention to context to recognize numbers with infinite sigfigs. This happens for one of two reasons. One is that the number measures something discrete rather than continuous: a number of objects, for example, where fractional values don't make sense. The other is defined values. For example, a centimeter is defined as one one-hundredth of a meter, so the conversion ratio between the two has infinite significant figures. This is true of most other conversion ratios between units of measurement. Similarly, the speed of light in a vacuum is, by definition, exactly 299,792,458 meters per second.

- The number of sigfig in a number is generally the number of digits in it, with the exception that trailing zeros don't count in numbers with no decimal point and leading zeros never count. (That is, the 0's are significant in "100." and "1.00", but not in "100" or "0.01". And, in "0.10", the first zero is insignificant but the second zero is significant.)

- When you need to look up values from the book or another reference, for example for fundamental constants, always use one or two more significant figures than the numbers in the problem. You don't want the number you looked up to constrain the number of significant figures in your result, but you also don't want to waste your time with a large number of digits that won't contribute more significant figures to the result.

- In addition and subtraction, a digit is significant only if it's the sum or difference of two digits that are significant or of a significant digit with a leading zero. So, in adding "123" and "45.67" to get "168.67", the first three digits of the result are significant but the last two are not, since they're the sum of significant ".67" with non-significant implied ".00".

- In multiplication and division, the answer has the fewer number of sigfigs of the numbers of sigfigs in the two numbers you've multiplied or divided.

- In a logarithm, the numbers to the right of the decimal point is called the mantissa and the number of significant figures must be the same as the number of digits in the mantissa. When taking anti-logarithms, the resulting number should have as many significant figures as the mantissa in the logarithm.[2]

- Once you have calculated your answer and determined the number of significant figures it contains, you just have to be sure to only write your answer with that many significant figures, rounding as necessary.

[2] Source: `http://en.wikipedia.org/wiki/Significant_figures`

3.2 Scientific and Engineering Notation

3.2.1 Scientific Notation

In physics, we often deal with numbers many orders of magnitude larger or smaller than one. However, we usually only care about a few significant digits in the numbers: when we say that earth is 149000000000 meters from the sun, we only really care about the first three digits, which are significant. However, decimal notation forces us to follow them by a large number of zeros and then count the total number of digits to determine the order of magnitude.

We can get around this problem by taking advantage of the fact that each place-value in the decimal system represents a power of ten: 456.78 is 4 hundreds, 5 tens, 6 ones, 7 tenths, and 8 hundredths. As a result, we can shift the position of the decimal point by multiplying by a power of ten: $4.5678 \times 10^2 = 456.78$. Numbers written in this format, with exactly one non-zero digit before the decimal place and followed by $\times 10^n$ where n is some integer, are said to be in "scientific notation."

Although numbers written in this form look like multiplication problems, they are considered to be single numbers, and they shouldn't be separated or treated differently. Furthermore, even though 10^1 is just equal to ten and 10^0 is just equal to one, it is standard to use these exponents in scientific notation just like any other exponents.

The utility of this system should be obvious: it allows us to write the distance from the earth to the sun as 1.49×10^{11} meters, and see immediately that it is larger than the 1.43×10^12 meters between Saturn and the sun, which would otherwise require us to carefully count the zeros in these very long numbers.

Using scientific notation allows physicists to consistently use a single unit for a given physical quantity instead of various units of different sizes. Instead of using inches for measuring clothing and miles for the distances between cities, we can pick a single length unit and use scientific notation to deal with lengths much larger or smaller than that unit.

Scientific notation makes it much easier to keep track of significant figures, since only significant digits are expressed: there is no reason to include the zeros in 3300 or 0.01 when writing them as 3.3×10^3 or 1×10^{-2}. Furthermore, it makes it possible to indicate whether 3300 has two or three significant figures. If it had four, we could write it as 3300., but decimal notation makes it impossible to indicate that some but not all of the trailing zeros are significant. On the other hand, 3.3×10^3 clearly has two significant figures while 3.30×10^3 has three.

Remember, if a number is given in scientific notation, all of the digits before the multiplication sign are significant, even if they are trailing zeros. However, the 10 and exponent are not significant.

3.2.2 Converting between Scientific and Decimal Notation

To convert a number from scientific notation to decimal notation, we could in theory simply perform the arithmetic operations indicated by the notation. However, there is a simpler method. We simply have to move the decimal point a number of places to the right or left as indicated by the exponent. If the exponent is positive, we should move it that many places to the right, and if it is negative, we should move it that many places to the left.

For example, 1.56×10^5 has an exponent of positive five, so we have to move the decimal point five places to the right: 156000. Meanwhile, 1.56×10^{-4} has an exponent of negative four, so we have to move the decimal point four spaces to the left: 0.000156.

Converting back from decimal notation to scientific notation just requires reversing this process. If we have to move the decimal point a certain number of spaces to the left to only have one digit before the decimal point, the exponent will be that number of spaces. If we have to move it to the right, though, the exponent will be the negative number of spaces.

For example, 0.045 requires us to move the decimal point two spaces to the right, so it is 4.5×10^{-2}. Meanwhile, 45 requires us to move it one space to the left, so it is 4.5×10^1.

It can be difficult at times to keep these rules straight. However, there is an easy way to determine whether one is moving the decimal point in the right direction for the exponent. If the exponent is positive, then the multiplication implied by scientific notation would make the number bigger; if the exponent is negative, then the multiplication would make the number smaller. Make sure that the direction you're moving the decimal point is consistent with this.

3.2.3 Engineering Notation

One variant of scientific notation that is not used in physics, but is relatively common in some other fields, is what is called "engineering notation." Engineering notation is very similar to scientific notation, except that the exponent to which ten is raised must be a multiple of three, and up to three digits are allowed before the decimal point.

Thus, 1.00×10^3 and 1.00×10^6 are acceptable in both scientific and engineering notation, but 1.00×10^4 would be written 10.0×10^3 and 1.00×10^5 would be written 100×10^3 in engineering notation.

3.3 The SI System of Measurements

3.3.1 Background on the SI System

Although Americans rarely use it, the metric system of measurement—meters for length, liters for volume, grams for mass, and °C for temperature—is nearly universal in the rest of the world. This near-universality makes it appealing to physicists, who want their results to be easily understood by physicists from other countries. However, like the US customary system of measurements, it was primarily designed for merchants to be able to measure volumes, weights, and lengths of their wares in a consistent manner.

In order to use it for scientific purposes, scientists have extended the metric system with new units to measure things like energy, electric current, force, and pressure. Unfortunately, they haven't found only one way to do so: there are two competing extensions of the metric system: "MKS units" and "cgs units." In each case

MKS units are by far the most common, and are the basis for the SI system of measurements, which is regulated by an international organization. Some fields in physics, especially those that study electricity and magnetism, prefer cgs units, however, so it is good to be aware that they exist.[13]

Like the metric system, the SI system uses a system of prefixes indicating powers of ten before unit names. These prefixes are equivalent to the suffixes of scientific notation, and can be replaced by them:

$$1\ \text{cm} = 1 \times 10^{-2}\ \text{m}$$

for example, since the prefix c, or centi, means 10^{-2}. If a unit is squared, the prefix is included in the square, thus, for example, for volume measured in cubic centimeters,

$$1\ \text{cm}^3 = 1 \times (10^{-2}\ \text{m})^3 = 1 \times 10^{-6}\ \text{m}^3$$

While we will rarely use these prefixes in this class—scientific notation is usually simple—they will come up from time to time and it is important to be familiar with them. A table of all the prefixes is given in the next section. The ones we will use most often in this class are centi (c, 10^{-2}), milli (m, 10^{-3}) and kilo (k, 10^3).

[13] You can recognize the use of cgs units in mechanics problems if mass is measured in grams, length in centimeters, time in seconds, energy in ergs, force in dynes, and so on.

3.3.2 Table of SI Prefixes

The following prefixes can be appended to the front of the names of base and derived base Si units (meter, second, Newton, etc). Likewise, their symbols can be appended to the fronts of the symbols for these units. They can't be stacked, however: kilometers and megameters are valid units, but not kilokilometers. Also, note that we append prefixes to gram as though it was the primary unit, rather than stacking them on kilogram: 10^3 kilograms is a megagram, not a kilokilogram.

Prefix	**Prefix Symbol**	**Scientific Notation**
yotta	Y	$\times 10^{24}$
zetta	Z	$\times 10^{21}$
exa	E	$\times 10^{18}$
peta	P	$\times 10^{15}$
tera	T	$\times 10^{12}$
giga	G	$\times 10^{9}$
mega	M	$\times 10^{6}$
kilo	k	$\times 10^{3}$
hecto	h	$\times 10^{2}$
deca	da	$\times 10^{1}$
deci	d	$\times 10^{-1}$
centi	c	$\times 10^{-2}$
milli	m	$\times 10^{-3}$
micro	μ	$\times 10^{-6}$
nano	n	$\times 10^{-9}$
pico	p	$\times 10^{-12}$
femto	f	$\times 10^{-15}$
atto	a	$\times 10^{-18}$
zepto	z	$\times 10^{-21}$
yocto	y	$\times 10^{-24}$

3.3.3 Fundamental and Derived SI Units

In order to make use of the SI prefixes, of course, we need units to append them to. The SI system is based on seven "base" units, which are defined based on physical constants or specific physical objects.[14] These units are the meter (m, length), kilogram (kg, mass), second (s, time), kelvin (K, temperature), mole (mol, amount of substance, used in chemistry), ampere (A, electric current), and candela (cd, "luminous intensity").

Only the first three of these units, the meter, kilogram, and second, will be used in this class. However, they're clearly not sufficient to measure everything we need to measure to talk about moving objects and their motion, nor about many other scientific quantities. As a result, "derived" metric units exist, which are defined in terms of the base metric units. Simple examples of this include measuring area in square meters (m^2), volume in cubic meters (m^3), and speed in meters per second (m/s or $\text{m}\cdot\text{s}^{-1}$). Metric prefixes can be used with these as well, and are always affixed to the first unit in the unit name. Thus, 10^3 m/s is one km/s, not one km/ks or k(m/s) or m/ms.

In theory, all derived quantities, including more complicated ones like energy or pressure, could be referred to in this way, and it is not incorrect to do so. However, since people are lazy and don't like to keep track of long bunches of units, some derived units have been given names, generally the names of famous scientists. Particularly relevant to us are the unit of force, the newton ($1\ \text{N} = 1\ \text{kg}\cdot\text{m}\cdot\text{s}^{-2}$); the unit of energy, the joule ($1\ \text{J} = 1\ \text{kg}\cdot\text{m}^2\cdot\text{s}^{-2}$); and the unit of pressure, the pascal ($1\ \text{Pa} = 1\ \text{kg}\cdot\text{m}^{-1}\cdot\text{s}^{-2}$). The radian is also treated as an SI derived unit, although since it is length divided by length, its fundamental unit components cancel. The liter (L) is also technically a derived unit, but it is defined as $1\ \text{dm}^3$, which means that $1\ \text{L} = 10^{-3}\ \text{m}^3$.

In order to check whether your answers make sense, it is useful to have a general sense of the size of the SI units in mind, although you'll need exact conversion ratios to do unit conversions. In length, one meter is about three feet and one centimeter is about two-and-a-half inches. In volume, one liter is just more than one quart. In mass, a kilogram is a bit over two pounds. And in speed, one meter per second is a bit over two miles per hour.

[14] The units were originally all based on physical standards kept in Paris, but all except the kilogram have been redefined in terms of experimental physical constants. There are plans to replace the kilogram with experimental measurement of physical constants in the next few years as well.

3.4 Unit Conversions

Converting between different units for the same quantity is necessary quite often in physics. Even if everyone used the SI system, some numbers might be written in kilometers and others in centimeters. However, many numbers are measured in non-SI units as well.

Unfortunately, converting between different units is also a very common source of error. In particular, people often multiply by a conversion ratio when they should divide by it, or vice versa. However, there is a very simple method for avoiding these errors.

Every conversion ratio can be thought of as a fraction that equals one. Since 1 ft = 12 in, $\frac{1\text{ ft}}{12\text{ in}} = 1$. Furthermore, since $1^{-1} = 1$, we can take the reciprocal of it without changing its value: $\frac{1\text{ ft}}{12\text{ in}} = \frac{12\text{ in}}{1\text{ ft}} = 1$. This is useful because we can multiply 1 by any expression without changing the expression's value. Thus, multiplying a number by a ratio of units that equals one will not change its value. However, it may allow us to cancel out units, resulting in a value in the desired units:

$$(24\text{ in})(\frac{1\text{ ft}}{12\text{ in}}) = \frac{24\text{ in}\cdot\text{ft}}{12\text{ in}} = 2\text{ ft}$$

With this method, it will be very clear to us if we have selected the wrong conversion ratio:

$$(24\text{ in})(\frac{12\text{ in}}{1\text{ ft}}) = \frac{(24)(12)\text{ in}\cdot\text{in}}{\text{ft}}$$

The units won't cancel properly and the result is nonsense.

3.5 Dimensional Analysis

3.5.1 The Important of Using Units

As you've probably noticed by now, science classes tend to have a lot of math in them. (After all, we just spent half of this class working learning math.) That said, there is one big difference between the math that we do in physics—or biology or chemistry—and the math that is done in math classes.

In physics, *every number has a unit attached.* Mathematicians tend to like to work with pure numbers: quantities that exist without being quantities of any specific thing. When they say that the roots of $y = x^2 - 1$ are ± 1, they don't mean $+1$ oranges -1 dollars, they just mean the numbers unattached to anything in the real world.

However, physics is concerned with things that we can actually measure, and you can't measure pure numbers. Any number you observe in a measurement has to have units—5 atoms, 20 grams, 40 cakes, 5 miles—rather than just being a bare number—5, 20, 4 tens, 5. Without units, you haven't actually communicated anything: if you tell me that the velocity of a rocket is "5", I don't know if you mean 5 mph, 5 m/s, 5% of the speed of light, or even if you've gotten confused and found that it's 5 kg.

If your final answer doesn't have units, it will most likely be counted as wrong. I won't guess what units you mean, even if the number you gave would be correct in commonly-used units.

Writing down the units in your final answer is essential to having your answer mean anything. However, it's a very good idea to write down your units at each step of the problem to keep track of them. This may seem annoying, but it makes it easier to figure out what what you were doing when you go back to look at your work later.

Furthermore, it makes it possible for us to use "dimensional analysis", a very useful technique for making sure that our answers and the formulas we use to get them make sense. Dimensional analysis allows us to determine what the units of an expression containing terms with different units in it should be. It also allows us to recognize nonsensical operations, such as adding five feet to ten pounds, that don't mean anything because they combine quantities that it doesn't make physical sense to combine.

3.5.2 Dimensional Analysis

Dimensional analysis is the practice of keeping track of the units of all numbers in an expression and combining them to determine what the units of the result of the expression will be. Doing so can be useful for figuring out the units of expressions that aren't in standard units, and also lets you determine whether an equation you have memorized can be correct.

There are several basic principles behind dimensional analysis:

- Every unit of measure can be expressed in terms of a few basic units of measure, called "dimensions". In this class, we will primarily care about units that are made up of three dimensions: time, length, and mass. For example, 1 m has a dimension of length only. 1 m/s and 1 mph both have dimensions of length divided by time.
- Numbers can only be added or subtracted if they have units of the same dimensions. To do this, they need to first be converted to the same units. Thus, $1.00\text{ m/s}+1.00\text{ mph} = 1.00\text{ m/s} + 0.45\text{ m/s} = 1.45\text{ m/s}$.
- Any numbers can be multiplied or divided or raised to an exponent. The units of the result can be determined by treating the operation as though the units were variable coefficients. Thus $5\text{ ft} \times 5\text{ ft} = 10\text{ft}^2$. Likewise, $(10\text{ m})/(5\text{ s}) = 2\text{ m/s}$.
- Finally, exponents themselves, as well as numbers that are the arguments of logarithms or inverse trig functions must be "unitless": their dimensions and units must cancel out, for example s/s or m/m. Meanwhile, of course, the arguments of trig functions must be angles.

As an example of these rules, consider the formula for kinetic energy, $E = \frac{1}{2}mv^2$, where m is mass and v is velocity. Since we know that mass has the dimension of mass and velocity has the dimensions of length divided by time, we can see that energy must have dimensions of mass times length squared divided by time squared. Furthermore, any other expression for energy must have the same dimensions to be valid. For example, Einstein's famous expression, $E = mc^2$, where E is energy, m is mass, and c is the speed of light in a vacuum has the same dimensions, even though it calculates a completely different sort of energy under different conditions.

Also, for example, if we measure mass in kilograms and velocity in meters per second, this expression will give energy in $\text{kg}\cdot\text{m}^2/\text{s}^2$. This combination of units is referred to with the SI derived unit of joules, and any equation that is supposed to give us energy in joules should produce the same combination of units.

While this can be quite useful, it's important to remember that it doesn't give us an exact form of the equation, just the dimensions. For example, it doesn't allow us to determine that kinetic energy is $E = \frac{1}{2}mv^2$ and not $E = 2mv^2$ or simply $E = mv^2$.

4 Forces and Motion

4.1 Displacement, Velocity, and Acceleration

4.1.1 Displacement versus Distance

One of the first—arguably the first—field of physics to be developed was the study of the behavior of moving bodies, called kinematics. The reason it was developed so early is that for many simple systems, such as the motion of projectiles (such as falling objects and cannonballs) and low-friction motion (such as wheeled vehicles), it is possible to describe the motion of objects in terms of only a few variables.

In order to begin studying kinematics (and thus physics), we first need to define these variables. They will be similar to the terms that we use in everyday language to describe motion, but not identical. In particular, since most motion occurs in more than one dimension and, even when it doesn't, direction matters, we will want to use vectors to describe the behavior of moving bodies.

The first everyday term for describing motion that one might think of is "distance" (d), i.e. how far an object has travelled. One can think of distance as the reading on an odometer: it increases with the length of our trip, and never goes back down even if we backtrack. If you drive twenty miles to Baltimore and twenty miles back, you've driven forty miles, not zero miles, even though you've begun and ended your trip in the same spot.

The problem with describing motion in terms of distance is that distance doesn't take into account direction, but it does take into account the path followed, the exact opposite of what a vector quantity should do. Instead, physicists usually prefer to study "displacement" ($\Delta\vec{x}$), the vector from where you started your trip to where you ended it. If you drive 20 miles north to Baltimore, your displacement is 20 miles north or (0 miles, 20 miles) if we consider north to be the $+x$ direction or (20 miles, 90°) if we set 0° as east. If you then return to where you started, though, your total displacement for the trip is 0, since the vector from where you began to where you ended is zero.

It's worth noting that this definition assumes that position itself can be defined as a vector, which makes sense if you think of plotting locations on a coordinate grid. Each position is described by an ordered pair, which is essentially the vector between it and the origin. Thus, we can define displacement as

$$\Delta\vec{x} = \vec{x}_1 - \vec{x}_0 \qquad \text{(Definition of Displacement)}$$

where $\vec{x_0}$ is the position at the beginning of the trip and $\vec{x_1}$ is the position at the end of the trip. (The notation of using Δ to mean "change in" and subscripts 0 and 1 to indicate final and initial values of a quantity is standard in physics and we will see a lot more of it.)

4.1.2 Velocity versus Speed

Speed is another concept that, like distance, we all have an intuitive understanding of: it's simply how fast you are going. There are really two types of speed one can measure, though: "instantaneous speed" (s_{inst}) and "average speed" (s_{ave}). Instantaneous speed is how fast you are going at a particular point in time; what your speedometer reads. Average speed, on the other hand, is the distance you've travelled divided by the time it took you to do it in, which you might get by timing your trip and using your odometer to see how far you went. We can describe average speed mathematically as

$$s_{ave} = \frac{d}{t_1 - t_0} = \frac{d}{\Delta t} \qquad \text{(Definition of Average Speed)}$$

where d is distance travelled and t_0 and t_1 are the times when the trip began and ended. Instantaneous speed is defined similarly, but with a value of Δt chosen that is small enough that the speed doesn't change at all during the interval.

The vector equivalent of speed is "velocity". Like speed, it comes in instantaneous and average forms. However, it is defined in terms of displacement divided by time. The "average velocity" ($\vec{v}_{ave}$) for a trip is the total displacement for the trip divided by the time the trip takes:

$$\vec{v}_{ave} = \frac{\vec{x}_1 - \vec{x}_0}{t_1 - t_0} = \frac{\Delta\vec{x}}{\Delta t} \qquad \text{(Definition of Average Velocity)}$$

where $\Delta\vec{x}$ is the total displacement for the trip and Δt is the amount of time the trip took. As with speed, the "instantaneous velocity" ($\vec{v}_{inst}$) can be found by taking this measurement with a value of Δt small enough that the velocity doesn't change at all during this interval.

Since instantaneous measurements of speed and velocity are taken so quickly that the quantities don't change during the measurement, instantaneous speed is just the magnitude of instantaneous velocity. Alternatively, instantaneous velocity is instantaneous speed plus a direction (10 mph north rather than 10 mph).

Average velocity is more than just average speed with a direction, however, because motions in different directions will have displacements that partially or completely cancel out. If you drive twenty miles north with a constant speed of 65 mph and then immediately change direction and drive twenty miles south with the same constant speed, your average speed is 65 mph: the direction you're traveling doesn't change the distance you travelled. However, since you returned to where you started, your total displacement is zero, and your average velocity is zero, even though your instantaneous velocity was never zero.

In general, the word velocity and the variables v, v_0, and v_1 will refer to instantaneous velocity. When we want to talk about average velocity, we will refer to it as average velocity and use the variable v_{ave}.

4.1.3 Acceleration

In the same way that velocity is the change of position—the displacement—over time, "acceleration" is the change in velocity over time. As with velocity, we can define both "average acceleration" ($\vec{a}_{ave}$) and "instantaneous acceleration" ($\vec{a}_{inst}$). Average acceleration is

$$\vec{a}_{ave} = \frac{\vec{v}_1 - \vec{v}_0}{t_1 - t_0} = \frac{\Delta\vec{v}}{\Delta t} \qquad \text{(Definition of Average Acceleration)}$$

where $\Delta\vec{v}$ is the change in velocity and Δt is the amount of time it took for the velocity to change. Instantaneous acceleration can be defined in the same way if we make Δt small enough that the acceleration doesn't change during the time interval.

In practice, acceleration can vary with time just as much as velocity can. During a long car trip, you may spend most of your time at constant velocity on a straight highway, but fairly frequently hit the gas pedal to increase your velocity or the brake to decrease it (such negative acceleration is sometimes called "deceleration").

However, studying systems with changing acceleration requires the use of calculus—actually, Newton's development of calculus was primarily motivated by his need for a mathematical framework for his studies of motion—so in this class we will limit ourselves to situations where acceleration is constant. As a result, average and instantaneous acceleration will always be the same, and we will simply refer to acceleration.

It is important to remember that velocity and acceleration are both vectors, and so have directions as well as magnitudes. Since velocity is a vector, a change in direction is a change in velocity, even if your speed never changes. Thus, a car going around a curve is accelerating, even if its speed is constant the whole time. This fact becomes especially important in understanding circular motion and why objects moving in circles feel a force pulling them outward.

4.1.4 A Note on Dimensions and Units

It is worth taking a moment to discuss the dimensions of and units used for the quantities we have just introduced. Displacement, like distance, is a change in position, so it has dimensions of length and the standard SI unit used for it is meters (m). Velocity and speed are both lengths divided by a change in time, so they have dimensions of length divided by time and the standard SI unit used for it is meters per second (m/s). Acceleration is velocity divided by time, so it has dimensions of length divided by time divided by time, i.e. length divided by time squared, and the standard SI unit used for it is meters per second squared (m/s^2).

4.1.5 Frames of Reference

An important concept that has been lurking behind the surface in our discussion of displacement, velocity, and acceleration is that of a "frame of reference." Our original introduction to displacement noted that position itself was defined as a vector from the origin of a coordinate grid to a point on that grid. We said nothing further about this grid, although all of our further definitions referenced position indirectly, and thus required the presence of such a grid.

When we are sitting on Earth's surface, it seems obvious that we should fix this grid to Earth's surface as well, defining some specific location as the origin and some specific direction as the $+x$ direction. However, this is overly restrictive. It turns out that the laws of physics will behave exactly the same in *any* frame of reference as long as it is not accelerating. If we are riding on an airplane or train that is moving very smoothly at a constant velocity, we can play catch or pinball exactly as we would standing on the ground. However, as soon as the vehicle hits a bump or some turbulence or accelerates, our game will be thrown off.

Obviously, the position and velocity of a thrown ball will be different if we measure them using a coordinate grid traveling at hundreds of miles an hour than if we measure them using a coordinate grid fixed to the ground. Thus, when we give a velocity or position, we need to specify them relative to a specific grid, which we call the frame of reference.

Of course, even though a passenger on an airplane at thirty thousand feet can describe the motion of a ball they throw down the aisle exactly as though the plane is sitting on the ground and not moving, the ball itself doesn't know that the airplane is there. From the point of view of an observer on the ground, it should follow the same laws of physics as a ball thrown by someone on the ground. This means that our choice of frame of reference is arbitrary: we can select whichever one is most convenient for solving a problem, usually one where the most objects are at rest.

The fact that we can choose which frame of reference to use implies that we can convert our description of an object's motion (its position, velocity, and acceleration) between frames of reference. Acceleration is simplest. Since frames of reference are only equivalent if they aren't accelerating, the acceleration of an object will be the same in all equivalent frames of reference.

If we want to convert position and velocity to a different frame of reference, we simply need to add the vector describing the difference in velocities or positions between the frames of reference to the velocity or position measured in the original frame of reference. For example, suppose we are walking at 3 mph towards the front of a bus that is driving 55 mph north. In the bus frame of reference, we are moving 3 mph forward (north). If we add this to the motion of the bus in the Earth's frame of reference, 55 mph north, we get a motion of 58 mph north in the Earth's frame of reference. On the other hand, if we are walking 3 mph towards the back of the bus, we'll add 3 mph south to 55 mph north and get 52 mph north in the Earth's frame of reference.

4.2 Motion Under Constant Acceleration

4.2.1 The Equations of Motion

Given that their definitions depend on each other, it should be clear that there is a relationship between time, position, displacement, velocity, and acceleration. For some simple problems, the relationships that are explicit in the definitions are sufficient for us to be able to find the value of one from the others. In other cases, though, it is useful to have additional equations relating them.

The following set of equations, known as the "kinematics equations", can be derived from the definitions of displacement, velocity, and acceleration through the use of calculus. In this class, however, we will simply accept them as given. All of them depend on the assumption that acceleration is constant over the period of time being considered.

$$\Delta\vec{x} = \vec{v_0}\Delta t + \frac{1}{2}\vec{a}(\Delta t)^2$$
$$\Delta\vec{x} = \vec{v_1}\Delta t - \frac{1}{2}\vec{a}(\Delta t)^2$$
$$\Delta\vec{v} = \vec{a}\Delta t$$
$$\vec{v_1}^2 - \vec{v_0}^2 = 2\vec{a}\cdot\Delta\vec{x}$$

In the kinematics equations, $\Delta t = t_1 - t_0$ is the time over which a motion occurs, $\Delta\vec{x} = \vec{x_1} - \vec{x_0}$ is the displacement from the beginning to end of the motions, $\Delta\vec{v} = \vec{v_1} - \vec{v_0}$ is the change in velocity from the beginning to the end of the motion, and $\vec{a}$ is the constant acceleration during the motion.

While time is a scalar, position, velocity, and acceleration are all vectors in the general forms of these equations. (Recall that a vector squared is just its magnitude squared, and note the dot product in the fourth equation.) However, the components of a vector are independent under addition, multiplication by a scalar, and the taking of dot products—the three operations performed here—which means that we can treat the x, y, and z components of motion completely separately by replacing the vector quantities with their components along a particular axis.

In simple cases, we may only need to solve the motion along a single axis: we may, for example, want to know how hard a dropped object will hit the ground, but not care exactly where it hits it. In those cases, we can simply solve the problem using the components along that axis and discard the other components. In more complicated situations, though, we may need to solve the motion along the two axes as though they are separate problems, linked only by the fact that the value of Δt will be the same for both problems, since time is a scalar and doesn't have separate components along different axes.

4.2.2 Gravity Near Earth's Surface

As we will see in the following examples, there are a number of situations that we can model with constant acceleration. One particularly interesting one is objects in free fall. We say an object is in "free fall" if the only vertical force on it comes from gravity. In other words, if there's nothing supporting it or pushing down on it, and there's no air resistance: the conditions in which Galileo proved that objects all fall at the "same speed."

More specifically, objects in free fall have a constant acceleration of 9.8 m/s^2 (32 ft/s^2) towards the ground, regardless of their mass.[15] This means that two objects dropped at the same time will have the same velocity at all times until they reach the ground, and if they are dropped from the same height, they will hit the ground at the same time. Because the acceleration of objects in free fall is used so frequently, it is given its own variable, $g = 9.6\ \text{m/s}^2$.

Three things are worth noting about this. First of all, objects in free fall only accelerate downward with an acceleration of $g = 9.6\ \text{m/s}^2$ near the Earth's surface. As we will see, the acceleration due to gravity depends on the mass of the object a falling object is falling towards and the distance between the two objects, so g decreases as we move further away from the Earth's surface and is quite different on other planets. However, since the earth is so large, a difference in altitude of a few miles results in a minimal change in g and it is reasonable to use $g = 9.6\ \text{m/s}^2$ even for objects dropped from airplanes.

Second of all, while objects only accelerate towards the ground with an acceleration of g when they are in free fall—when there are no other vertical forces on them, such as air resistance, an initial vertical velocity does not change their vertical acceleration. An object thrown towards the ground or thrown up into the air will still have an acceleration towards the ground of g, although its velocity as a function of time will of course be different.

Third of all, recall that the horizontal and vertical components of the acceleration, velocity, and position vectors are independent. This means that whether an object is being propelled horizontally is irrelevant to whether it is in free fall, and that two objects with different horizontal motion will still accelerate towards the ground at the same rate. For example, imagine two marbles dropped off a table at the same time. One of them was held stationary at the edge of the table before being dropped, the other was rolling at high speed across the table before falling off. Although they will land in very different places, both will hit the ground at exactly the same time.

[15] Although this acceleration can be measured much more precisely than to two significant figures, it varies over the surface of the Earth due to the fact that the Earth isn't a perfect sphere, and it is only constant over the Earth's surface to two significant figures.

4.2.3 Using the Equations of Motion in One Dimension

We can use the kinematics equations to solve a variety of problems involving motion in one dimension under constant acceleration. To do this, we generally want to select an equation that involves the quantity we are looking to find and quantities that we know, but no quantities that we don't know. We then rearrange this equation to solve for the quantity of interest. The best way to learn to do this is through experience. To that end, several examples of one-dimensional kinematics problems follow.

Example: Measuring the Depth of a Well

Suppose you want to measure the depth of a well. You drop a coin in the well and, 0.85 s later, you hear a splash. What is the depth of the well?

We know that Δt is 0.85 s, and that v_0 is 0.0 m/s (since we dropped the coin from rest rather than throwing it). Since the object is in free fall—a coin is small and dense, so air resistance is negligible—we know that $a = g$, that is 9.8 m/s down. We want to find Δx. Conveniently, there is one equation that contains exactly these variables:

$$\Delta x = v_0 \Delta t + \frac{1}{2} a (\Delta t)^2$$

Let's define up as positive and substitute numbers into the equation:

$$\Delta x = (0.0 \text{ m/s})(0.85 \text{ s}) + \frac{1}{2}(-9.8 \text{ m/s}^2)(0.85 \text{ s})^2$$

We can simplify this to

$$\Delta x = 0.0 \text{ m} + \frac{1}{2}(-9.8 \text{ m/s}^2)(0.7225 \text{ s}^2)$$

which evaluates to

$$\Delta x = -3.54025 \text{ m}$$

This number is negative since the coin moved downward, in the $-x$ direction, and has two significant figures since the number given in the problem and the constant g had two significant figures. (We can assume that we know the initial velocity of zero to within two significant figures as well, since a tenth of a meter per second would be about a third of a foot per second, and we can presumably tell the difference between dropping a coin and throwing it that fast.)

Thus, the well is 3.5 m deep.

Example: Breaking Distance

A particular car's breaks can reduce its velocity by 15 m/s each second. If it takes 0.45 s for a person to slam on the breaks after they see an obstacle, how far will the car travel from the moment they see the obstacle until the car comes to a complete stop if it is initially traveling at 25 m/s?

This problem actually has two components. First, the car will travel at a constant velocity from the time the obstacle is sighted until the breaks are activated. Then, the car will decelerate with a constant acceleration until it reaches a final velocity of 0.0 m/s.

For the first problem, we can use

$$\Delta x = v_0 \Delta t + \frac{1}{2} a (\Delta t)^2$$

substituting in 0.45 s for Δt, 25 m/s for v_0, and 0.0 m/s^2 for acceleration:

$$\Delta x = (25 \text{ m/s})(0.45 \text{ s}) + \frac{1}{2}(0.0 \text{ m/s}^2)(0.45 \text{ s})^2$$

This evaluates to 11.25 m. We will refrain from rounding this down to its two significant figures until the problem is completed.

For the second problem, we know the initial and final velocities and the acceleration, but not the travel time, so we can use

$$v_1^2 - v_0^2 = 2a\Delta x$$

substituting in 25 m/s for v_0, 0.0 m/s for v_1, and -15 m/s^2 for a. The acceleration value is negative because the car is slowing down: its acceleration is in the opposite direction as its motion, which we've indicated is positive by the sign of v_0. This gives us

$$(0.0 \text{ m/s})^2 - (25 \text{ m/s})^2 = 2(-15 \text{ m/s}^2)\Delta x$$

which simplifies to

$$-625 \text{ m}^2/\text{s}^2 = -30. \text{ m/s}^2 \Delta x$$

Using algebra, we can solve for Δx:

$$\Delta x = \frac{-625 \text{ m}^2/\text{s}^2}{-30. \text{ m/s}^2} = 20.83 \text{ m}$$

The total forward motion of the car from the moment the driver sees the obstacle until the car reaches a complete stop will be the sum of these two distances: 11.25 m + 20.83 m = 32.08 m.

Since all of the values in the original problem had two significant figures, the two distances we added each had two significant figures. The first two digits in the sum were produced by the sum of these figures, so the final answer has two significant figures: 32 m.

Example: A Ball Thrown Vertically

Suppose you are standing on the edge of a 12 m tall cliff and you throw a ball upward with a velocity of 15 m/s. How long will it take for the ball to hit the ground at the bottom of the cliff?

To begin with, we know v_0, 15 m/s, and Δx, which is -12 m because the ball's final location will be 12 m below where it is initially thrown. Although the ball starts out with a non-zero initial velocity, no forces other than gravity act on it after it leaves your hand, so it is in free fall during its entire motion, with $a = g$, $-9.8\ \mathrm{m/s^2}$.

Since we know Δx, v_0, and a, we can use

$$\Delta x = v_0 \Delta t + \frac{1}{2} a (\Delta t)^2$$

to solve for Δt. We begin by substituting in the known values:

$$(-12\ \mathrm{m}) = (15\ \mathrm{m/s})\Delta t + \frac{1}{2}(-9.8\ \mathrm{m/s^2})(\Delta t)^2$$

Rearranging this as

$$0 = \frac{1}{2}(-9.8\ \mathrm{m/s^2})(\Delta t)^2 + (15\ \mathrm{m/s})\Delta t - (-12\ \mathrm{m})$$

gives us a quadratic expression in Δt. We can use the quadratic formula to solve it:

$$\Delta t = \frac{-15\ \mathrm{m/s} \pm \sqrt{(15\ \mathrm{m/s})^2 - 4(\frac{1}{2}(-9.8\ \mathrm{m/s^2})(12\ \mathrm{m})}}{2(\frac{1}{2}(-9.8\ \mathrm{m/s^2}))}$$

This gives us two values of Δt, -0.66 s and 3.7 s, both with two significant figures. Since time should be positive, we can discard the negative answer and conclude that the ball will hit the ground 3.7 s after it is thrown.[16]

[16] The negative root is due to the fact that the ball's trajectory passes through the ground at the bottom of the cliff twice. If you imagine that, instead of being thrown at $t = 0$ s, the ball had already been moving in free fall, it would have had to passed through the ground below the cliff 0.66 s before it reached your hand.

4.2.4 Using the Equations of Motion in Two Dimensions

The kinematics equations can also be used to solve problems involving motion in two dimensions. Since the x and y components of motion are completely independent, we can solve the equations for motion along one dimension while ignoring motion along the other dimension.

However, since Δt is a scalar, each component of the motion must take the same amount of time. This allows us to use knowledge about the motion in one dimension to learn things about the motion in the other direction, as will be seen in the following examples.

Example: The Range of a Cannon

A cannonball is fired with a velocity of 100. m/s at a $45°$ *angle above flat ground. How far away will it land?*

The information that we want—how far away the cannonball will land—can only be found by solving the kinematics equations for the cannonball's horizontal motion to find the horizontal value of Δx. However, while we know that the horizontal acceleration is 0.0 m/s and the initial and final horizontal velocities are 100cos(45°) m/s, this isn't enough information to find the horizontal displacement without knowing how long the cannonball is in the air.

To find this time, we first solve the vertical problem, using 100sin(45°) m/s for v_0, 0.0 m for Δx (because the cannonball will land at the same elevation it was launched at), and g, -9.8 m/s^2 for a.

$$\Delta x = v_0 \Delta t + \frac{1}{2} a (\Delta t)^2$$

$$0.0 \text{ m} = (70.7 \text{ m/s})\Delta t + \frac{1}{2}(-9.8 \text{ m/s}^2)(\Delta t)^2$$

Conveniently, we can solve this without the quadratic equation. A term of Δt can be factored out,[17] giving us

$$0.0 \text{ m} = (70.7 \text{ m/s}) + \frac{1}{2}(-9.8 \text{ m/s}^2)(\Delta t)$$

which we can solve for Δt to find out that the cannonball will be in the air for 14.4 seconds. We can then use this value of time in the horizontal problem, substituting the values we know into the same kinematics equation:

$$\Delta x = v_0 \Delta t + \frac{1}{2} a (\Delta t)^2 = (70.7 \text{ m/s})(14.4 \text{ s}) + \frac{1}{2}(0.0 \text{ m/s}^2)(14.4 \text{ s})^2$$

Evaluating this equation gives us a value of 1.0×10^3 m for Δx. The cannonball travels about one kilometer, nearly two-thirds of a mile.

[17] This term corresponds to the fact that the cannonball is also at ground level at t_0, when it is launched.

Example: A Package Dropped from an Airplane

An airplane flying horizontally with a velocity of 65 m/s drops a package from a height of 750 m. What is the package's velocity in magnitude-angle form when it hits the ground?

Since the package is not accelerating in the horizontal direction, the horizontal component of its velocity will still be 65 m/s when it strikes the ground. To find the vertical component of its velocity, we have to make use of the fact that its vertical acceleration is g, 9.8 m/s towards the ground. Since we know the distance it will fall (750 m) and its initial vertical velocity (0.0 m/s), we can use the equation

$$v_1^2 - v_0^2 = 2a\Delta x$$

substituting in these values and noting that acceleration and displacement will be negative, since it is falling downward.

$$v_1^2 - (0.0\text{ m/s})^2 = 2(-9.8\text{ m/s}^2)(-750\text{ m})$$

This can be solved for v_1 to find that the final velocity is -121 m/s. (Mathematically, both positive and negative 121 m/s are possible. The positive value would correspond to a situation where the signs of acceleration and displacement are both reversed, and the object started with zero velocity on the ground and was accelerated upward.)

Thus, in component form we have the final velocity as (65 m/s, -121 m/s). The magnitude will be $\sqrt{65^2 + 121^2} = 137$ m/s. The direction can be found by evaluating $\arctan(121/65) = 61.8°$ and noting that the vector will be in the fourth quadrant. Thus, the final velocity vector is 140 m/s and 61.8° below horizontal (or 298.2°).

4.3 Newton's Laws of Motion

4.3.1 The First and Second Laws: Defining Force

So far, we have been discussing acceleration as though it's just something that happens to objects: they have a given acceleration, and we do kinematics calculations based on this acceleration, but we haven't talked about why they do or don't accelerate in different circumstances.

Of course, we know from our everyday lives that objects accelerate because other objects push or pull on them. Normally, the objects pushing or pulling on them need to be in physical contact with them. However, gravity and magnets are special cases: falling objects are pulled towards the Earth even when they're not touching it, and magnets can attract or repel at a distance. However, even in this case, the objects need to be pushed or pulled on to accelerate.

At first glance, we might think that while positive accelerations require an object to be pushed or pulled, negative accelerations can happen spontaneously: that it's in the nature of an object to come to a rest if nothing is pushing on it. However, more careful experimentation shows that the tendency of objects to slow down in everyday life comes from air resistance and friction, which are both cases of an object being pushed on by its environment.

Furthermore, we know that how heavy an object is seems to correlate to how hard it is to accelerate it. A slight tap can make a tennis ball move at a significant speed, but it takes a lot of effort to get a bowling ball moving. Air resistance will slow the fall of a heavy object less than a light object of the same size.

Newton's first and second laws of motion are essentially attempts to quantify this behavior of matter. His **First Law of Motion**—actually a special case of the second law–is as follows: "When viewed in a non-accelerating reference frame, an object has zero acceleration unless it is acted upon by an outside force."

The more general **Second Law of Motion** quantifies what happens when an object *is* acted upon by an outside force:

$$\Sigma\vec{F} = m\vec{a}$$

where $\vec{a}$ is the acceleration of the object, m is its "mass," $\vec{F}$ is the "force" exerted on the object, and Σ is the "summation sign", indicating that we are to add all such forces.

There is quite a lot of new material in this equation for us to unpack. We have defined two new physical quantities, mass and force, as well as their relationship to one another and to acceleration.

Mass—which we already have an intuitive sense of, and are familiar with the units for, kilograms—is a scalar and is essentially a measure of how hard it is to make an object change its velocity. Mass is a fundamental "extensive property" of matter, which means that the total mass of an object is the sum of the masses of its components. It does not necessarily correspond to the size of an object, however, since some objects are denser than others: a small rock can be much more massive than an inflated balloon, although the balloon is much bigger.

The property of mass that makes an object resist acceleration is called "inertia." However, mass has a second interesting property unrelated to this one: the mass of an object is also proportional to the force gravity exerts on it. In other words, an object that is twice as massive will be twice as heavy.

For this reason, we frequently speak of the weight of an object and its mass as interchangeable, though the weight depends on the object being near Earth's surface and the mass does not. (Formally, kilograms are units of mass and should not be used to measure weight, while pounds are units of force, and should not be used to measure mass, but near Earth's surface, where the strength of gravity is roughly constant, it is common to do so.)

Force is a less intuitive concept, but one that we still have some sense of. It is a vector quantity measuring the push or pull on an object that causes the object to accelerate. It has the same direction as the acceleration it causes, and we can find its magnitude by measuring the acceleration it causes and multiplying that acceleration by the mass of the object.

Dimensional analysis gives us the units of force. Since it is the product of acceleration (measured in $\mathrm{m/s^2}$) and mass (measured in kg), we measure mass in $\mathrm{kg{\cdot}m/s^2}$. This SI derived unit is called the newton and given the symbol N.

In the same way that it only makes sense to talk about acceleration in the context of a specific object that is accelerating, it only makes sense to talk about force in the context of a specific object it's acting upon.

It is important to remember, however, that Newton's Second Law refers to the sum of all forces acting on an object. Since force is a vector quantity, we need to add these forces as vectors. If I am pushing on an object with 20 N to the right, and you are pushing on it with 20 N to the left, the total force on the object is 0 N and it won't move at all.

Now that we have a sense of what force is, we can see that the statement that the mass of an object is proportional to its weight is equivalent to the statement that all objects in free fall accelerate at the same constant rate.

By definition, the only force on an object in free fall is its weight force, which we can express as its mass times a constant vector, $\vec{g}$, so $\vec{F_w} = m\vec{g}$. Since the total force on an object—in this case, just its weight force—is equal to its mass times its acceleration, we have $\vec{F_w} = m\vec{g} = m\vec{a}$, so all objects in free fall have a constant acceleration of $\vec{g} = 9.8\ \mathrm{m/s}^2 = 9.8\ \mathrm{N/kg}$ downward.

4.3.2 Relating Acceleration and Forces

Newton's second law of motion allows us to determine the acceleration of an object if we know all of the forces on it: we simply need to add the forces (as vectors, of course) and divide the result by the object's mass. We can also rearrange it to solve other problems, however.

If we know the net force on an object and its acceleration, we can determine its mass by dividing the magnitude of the net force by the magnitude of the acceleration: this is how most scales work, in fact. And, if we know the acceleration and mass of an object, we can determine what the net force on it is. This may also allow us to establish the value of an unknown force, if all other forces on the object (or forces on the object along a certain dimension) are known.

Example: Finding Acceleration with the Second Law

An airplane with a mass of 2.5×10^3 *kg has engines producing 45 kN of thrust at an angle* 5° *above horizontal. If the lift force produced by the wings is 25 kN directly upward, what is the magnitude and direction of the acceleration of the aircraft?*

There are three forces acting on the aircraft: its weight force, $(2.5 \times 10^3 \text{ kg})(9.8 \text{ N/kg}) = 24500 \text{ N}$ downward; its lift force, $25 \text{ kN} = 25000 \text{ N}$ upward; and its thrust, $45 \text{ kN} = 45000 \text{ N}$ at an angle 5° above horizontal. We need to break the thrust force into horizontal and vertical components to add the three: the vertical component is $(45000 \text{ N})(\sin 5°) = 3922 \text{ N}$ upward and the horizontal component is $(45000 \text{ N})(\cos 5°) = 44829 \text{ N}$ forward.

Thus, in the vertical direction, the net force is $25000 \text{ N} + 3922 \text{ N} - 24500 \text{ N} = 4422 \text{ N}$ upward. In the horizontal direction, the net force is 44829 N forward. Thus, the vertical acceleration is $(4422 \text{ N})/(2500 \text{ kg}) = 1.77 \text{ m/s}^2$ upward and the horizontal acceleration is $(44829 \text{ N})/(2500 \text{ kg}) = 17.03 \text{ m/s}^2$ forward.

The magnitude of the acceleration can be found by the Pythagorean theorem, as isusual for vectors: $\sqrt{(17.03 \text{ m/s}^2)^2 + (1.77 \text{ m/s}^2)^2} = 17 \text{ m/s}^2$, to the two significant figures given in the problem.

The direction of the acceleration will be $\arctan(|1.77 \text{ m/s}^2|/|17.03 \text{ m/s}^2|) = 5.9°$ above horizontal. This is a slightly higher angle than the thrust force, because the lift exceeds the weight, adding an additional upward component to the vertical acceleration.

Example: Finding Mass with the Second Law

A rocket whose engine produces a thrust of 500 N is launched vertically. It is observed to accelerate at 15 m/s^2. What is its mass?

The rocket's net force in the vertical direction is the thrust minus the weight force: $(500\text{ N}) - (m)(9.8\text{ N/kg})$. This net force should equal the product of its acceleration and mass, giving us:

$$(500\text{ N}) - (m)(9.8\text{ N/kg}) = (m)(15\text{ m/s}^2)$$

We can rearrange this equation to solve for the mass:

$$(500\text{ N}) = (m)(15\text{ m/s}^2 + 9.8\text{ N/kg})$$

noting that N/kg and m/s^2 are equivalent units, since 1 N is 1 kg·m/s^2. Thus, the mass of the rocket is 20 kg (with the same one significant figure we were given in the problem for the thrust).

Example: Finding a Force with the Second Law

A helium balloon with a mass of 15 g has an acceleration upward of 0.5 m/s^2. What is the lift force on the balloon?

Two forces are acting on the balloon, both in the vertical direction: its weight force, $(0.015\text{ kg})(9.8\text{ N/kg}) = 0.147\text{ N}$ and the unknown lift force. Since the balloon's acceleration is 0.5 m/s^2 upward, the net force on it must be $(0.015\text{ kg})(0.5\text{ m/s}^2) = 0.0075\text{ N}$ upward.

This gives us the equation $F_{\text{lift}} - (0.147\text{ N}) = 0.0075\text{ N}$, which we can solve for F_{lift} to find that the lift force must be 0.1545 N upward. Since the problem only gave us one significant figure for the acceleration, this is roughly 0.2 N upward.

4.3.3 The Third Law: Reaction Forces

While Newton's first and second laws of motion involve the effects of forces on a single object, his third law tells us about the interaction between two forces. According to the **Third Law of Motion**, "When an object exerts a force on a second object, the second object simultaneously exerts a force on the first body that is equal in magnitude but opposite in direction from the first force."

This law can be a bit confusing. In particular, it is important to realize that the two "reaction force" does not act on the same body as the original force. If it did, the two forces would cancel out and acceleration would be impossible!

Instead, what the third law says is that if you exert a force on an object by pushing on it, it will push back with an equally-strong force. This is why guns recoil when fired, why objects bounce off surfaces, and why you need to get a firm footing before pushing a massive object. If you are standing on something slippery, like ice or a well-waxed floor, the reaction force pushing back on you may be enough to accelerate you more than the more-massive object you're trying to move. If you have a firm footing, though, you can use your feet to push on the ground in the opposite direction, and its reaction force on you will cancel the reaction force of the object you are trying to move.

4.3.4 Normal and Tension Forces

Two particular types of reactive forces are of particular interest to us, and they are given special names: normal forces and tension forces. A "normal force" is simply the force produced by an object pushing on a surface. It is so called because it is perpendicular to ("normal" is another word for perpendicular) to the surface you are pushing on. This is why a ball thrown at a surface at an angle will bounce away in a different direction, and it more generally means that slanted surfaces can be used to change the direction of forces.

The component of the normal force parallel to the original force that produces it has to be equal and opposite to that force to satisfy the third law of motion. However, since the total normal force is perpendicular to the surface, there must be a second component of the normal force perpendicular to the original force and of a magnitude that when added to the parallel reaction force produces a force perpendicular to the surface.

The concept of a "tension force" is rather simpler: it is simply the force exerted on a rope or chain when the rope or chain pulls on an object. The important thing to keep in mind about the tension force is that if both ends of the rope are pulling on things, the tension forces need to balance of the whole rope will move towards one or the other object.

Example: Weight on a Scale

A person with a mass of 55 kg is standing on a scale in an elevator. The scale measures the total force they exert on it, which must be equal to the normal force it exerts on them. What weight does the scale read when the elevator is stationary? When it is moving upward with a constant velocity of 2.5 m/s? When it is accelerating upward at 2.5 m/s^2?

The person's acceleration times their mass must equal the net force on them. Whether the elevator is stationary or moving at a constant velocity, their acceleration is zero, which means that the net force on them must be zero as well.

Since the person has only two forces on them, the weight force—$wg = (55\text{ kg})(9.8\text{ N/kg}) = 539\text{ N}$—and the normal force, these two forces must cancel exactly. Thus, the normal force is 539 N upward and the scale registers a weight of 540 N (to two significant figures).

However, if the elevator is accelerating upward at 2.5 m/s^2 then the person presumably is as well. This means that the net force on them must be $F = ma = (55\text{ kg})(2.5\text{ m/s}^2) = 137.5\text{ N}$ upward. Since their mass and thus the weight force is unchanged at 539 N downward, the normal force, which is the only other force acting on them, must be $539\text{ N}+137.5N = 676.5\text{ N}$ upward.

Since the normal force acting on the person from the scale is 676.5 N upward, the force they exert on the scale must be the same in the opposite direction. Thus, to two significant figures, their weight registers as 680 N.

This discovery, that our weight increases when we are accelerating upward and decreases when we are accelerating downward should not come as a surprise: we know that we feel heavy when an elevator in a tall building accelerates quickly, and that we feel weightless for a moment when we are in free fall when jumping. Likewise, we know that an accelerating car pushes us back into our seats, and a sudden stop (with the accompanying rapid deceleration) throws us forward.

However, the realization that the effects of gravity and acceleration are indistinguishable—that we cannot tell if the closed box we are in is sitting on the surface of a planet and experiencing gravity or in empty space but accelerating—is actually quite important. It was one of the basic realizations that led Einstein to his development of the theory of General Relativity, which integrates gravity into his earlier theory of Special Relativity, about acceleration and velocity at close to the speed of light.

Example: Pulleys

Consider the two blocks shown in Figure 4. What force needs to be supplied to each rope to keep the blocks from accelerating?

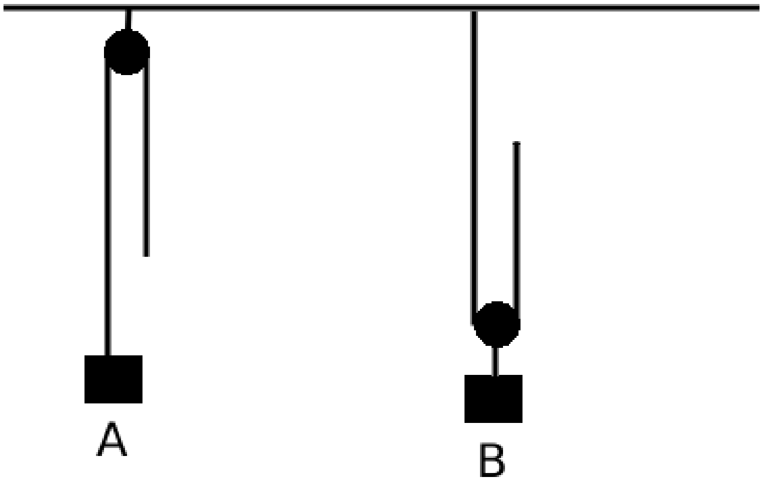

Figure 4: Blocks A and B are each supported by pulleys with one end of the rope free to be pulled by a person. Block A is attached to the other end of its pulley rope, with the pulley secured to the ceiling. Block B is attached to its pulley directly, with the other end of the rope attached to the ceiling. Both blocks have masses of 1.0 kg.

Two forces act on block A, its weight force, (1.0 kg)(9.8 N/kg) = 9.8 N downward and the force exerted by the rope upward. If the block is not accelerating, these two forces must be balanced, which means the rope is exerting a force of 9.8 N upward and the block is producing a tension of 9.8 N in the rope.

Since the tension must be the same on both ends of the rope for it to be taut without stretching, a downward force of 9.8 N is needed on the loose end of the rope to produce the same tension on this end and keep block A from accelerating.

Block B experiences the same weight force as block A. However, it is supported by two segments of the same rope. Since the tension is uniform throughout the rope, both segments must exert the same force on the block: half its weight force, or 4.9 N. Each end of the rope must be suspended with the same force upward to balance the tension, so a force of 4.9 N is needed on the loose end of the rope to support block B.

4.3.5 Dispersive Forces: Air Resistance

In everyday life, we are used to the idea that moving objects naturally come to a stop unless something is actively pushing on them. However, Newton's second law of motion tells us that this is *not* the natural behavior of moving objects. Instead, the fact that many objects slow down unless a force continuously acts to keep them moving is due to what are called "dispersive forces." Dispersive forces are called this because they take a moving object's kinetic energy—the energy of motion—and disperse it as heat, slowing the object down. They always produce an acceleration in the opposite direction as motion.

The two main dispersive forces that we come into regular contact with are drag and friction. "Drag" is the force that a fluid like air or water exerts to resist the motion of an object through it; for a falling object we often call it "air resistance." Calculating the drag exerted on an object is a complicated fluid dynamics problem that depends on the shape of the object, the properties of the fluid it's moving through, and the object's velocity.

What is consistent about drag, however, is that for a given object, it increases the faster the object is going. This means that, if an object is falling fast enough, the drag force on the object will eventually be equal to the weight force on the object. The net force on the object will be zero and so its velocity will remain constant until it hits the ground. The velocity at which this occurs is called "terminal velocity."

Terminal velocity is the reason that small animals, particularly insects, can fall from huge heights and be unharmed. Since their mass is very low, the weight force on them is as well, and they don't need to be moving very quickly for the drag force to cancel it out. Likewise, this is why raindrops don't kill us. A raindrop falling from a cloud several miles up would strike with a speed similar to a bullet if there was no air resistance. Instead, since it is very light, air resistance limits its speed to a rather low terminal velocity.

Example: Air Resistance and Terminal Velocity

A skydiver with a mass of 85 kg jumps out of a plane. What is the net vertical force on him immediately after he jumps out? What is the net vertical force on him when he reaches terminal velocity? When he opens his parachute, it significantly lowers his terminal velocity. Which direction is the net force on him at the moment the parachute opens?

When the skydiver first jumps out of the plane, he isn't moving in the vertical direction yet, so air resistance doesn't exert a force on him, and his net vertical force is $mg = (85\text{ kg})(-9.8\text{ N/kg}) = 830\text{ N}$.

When he reaches terminal velocity, by definition, the drag force and weight force cancel out, so the net force on him is 0 N.

When the parachute opens, it increases air resistance so that it exceeds the weight force and he is moving faster than his new terminal velocity. The net force on him is upward, reducing his velocity, until he reaches his new terminal velocity.

4.3.6 Dispersive Forces: Friction

"Friction," of course, is the force exerted by a surface on an object sliding across it. The strength of this force depends on the amount of surface area in contact with the surface and the nature of the surfaces—sandpaper produces more friction than ice—but we can summarize the magnitude of the force from friction as

$$\|F_{fr}\| = C_{fr}\|F_N\|$$

where $\|F_{fr}\|$ is the magnitude of the force from friction, $\|F_N\|$ is the magnitude of the normal force exerted by the surface, and C_{fr} is a constant called the coefficient of friction that depends on the specifics of the surfaces in contact. The direction of the friction force, unsurprisingly, is parallel to the surface and in the opposite direction as the motion of the object.

However, if an object is not moving, it can still experience a friction force, called "static friction." This force is still parallel to the surface, but is equal and opposite to the net force being exerted on the object along the surface. However, unlike "kinetic friction" (the friction of moving objects), there is a maximum possible static friction force, which is calculated in the same way as kinetic friction. If this force is exceeded, the object will start moving, and kinetic friction will take over.

Furthermore, the coefficient of friction is different for static and kinetic friction, and the coefficient of static friction is generally higher than the coefficient of kinetic friction. This means that a stationary object can resist a rather large force without moving, but once it starts to move, the friction force will suddenly be reduced. This is why stuck screws or lids that you are trying to loosen will often suddenly seem to break free: the breaking feeling is the switch from static to kinetic friction and an overall reduction in the friction.

Example: Friction and Motion

A block with a mass of 5.3 kg is sitting on a surface. The coefficient of kinetic friction is 2.1. If the block is sliding with a velocity of 12 m/s and there are no forces on it but friction, how long will it take for it to come to a stop?

Since the block masses 5.3 kg, the weight force it experiences is (5.3 kg)(9.8 N/kg) = 51.94 N. The friction force can be found by multiplying this by the coefficient of kinetic friction, (2.1)(51.94 N = 109.074 N). By Newton's second law of motion, the acceleration of the block is $a = F/m = 109.074$ N/5.3 kg = 20.58 m/s^2.

We can use the kinetics equation $\Delta v = a\Delta t$ to determine the time for the block to stop if we rearrange it to solve for Δt. The block's velocity will change from 12 m/s to 0 m/s in (12 m/s)/(20.58 m/s^2) = 0.58 s.

Example: Static and Kinetic Friction

Consider the situation illustrated in Figure 5. If the system is set up as shown with an initial velocity of 0.0 m/s, what will the acceleration of block B be? What will happen to this acceleration if block B is then tapped slightly, causing the system to begin to move?

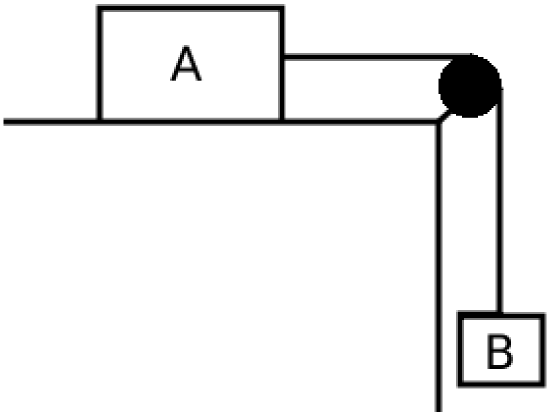

Figure 5: Block A has a mass of 4.0 kg, a coefficient of static friction of 0.55, and a coefficient of kinetic friction of 0.45. Block B has a mass of 2.0 kg. The pulley and rope are massless and frictionless.

Two forces act on block B: its weight force of (2.0 kg)(9.8 N/kg) = 19.6 N and the tension force from the rope. For the rope to be taut and not stretching, it must exert an equal force on block A, which can only be cancelled by block A's friction force.

Since block A has a coefficient of static friction of 0.55, its maximum static friction force is (0.55)(4.0 kg)(9.8 N/kg) = 21.56 N. This is greater than the force of 19.6 N to the right that the rope transmits to the block, so static friction provides a force of 19.6 N to the left, and neither box moves. (Both have an acceleration of 0.0 N.)

If block B is tapped, starting the system in motion, the weight force on it will remain the same. However, block A will be in motion, and thus in the kinetic friction regime, where the coefficient of friction is lower. The friction force exerted on block A will now be (0.45)(4.0 kg)(9.8 N/kg) = 17.64 N to the left and the net force on block B will be 19.6 N − 17.64 N = 2.0 N downward.

Since this net force is acting on the two-block system with a total mass of 6.0 kg, it will result in an acceleration of $\vec{F}/m$ = (2.0 N)/(6.0 kg) = 0.33 m/s^2

4.4 Circular Motion and Newton's Law of Universal Gravitation

4.4.1 Circular Motion: Centripetal Acceleration

So far, we have discussed forces and accelerations that have constant components. Although these forces may result in curved motion of the horizontal and vertical components of the acceleration differ, the horizontal and vertical components of the acceleration are each constant, allowing us to solve the kinematics equations in each dimension with the constant-acceleration kinematics equations.

Of course, not all motion has this property: acceleration often changes with time, and the components of it often change differently. Unfortunately, as noted before, solving problems with changing acceleration generally requires the use of calculus. However, there is one special exception to this: circular motion.

When an object moves in a circle at a constant speed—for example, a horse on a merry-go-round—it is constantly accelerating, since it is constantly changing direction. Furthermore, the acceleration is changing direction as well: the object's velocity is always curving towards the center of the circle, even though which direction the center of the circle is depends on the location of the object. (When it's at the northern extreme of its motion, the center of the circle is to the south, but when it's at the southern extreme, the center of the circle is to the north.)

However, while the direction of the acceleration is constantly changing, its magnitude is constant. Otherwise, the curvature of the circle wouldn't be constant, and the object would sometimes move closer to the center and sometimes further away. Furthermore, while the direction is constantly changing, the fact that it is constantly directed towards the center of the circle means that some problems can be solved easily. In particular, we can determine the force needed to keep the object moving circularly.

It turns out that it can be shown geometrically that the acceleration of a body moving in a circle at a constant speed—"called centripetal acceleration"—is

$$a_{centripetal} = \frac{v^2}{r}$$

towards the center of the circle where v is the velocity of the object at any point in time and r is the radius of the circle. (Since the square of a vector is simply the square of its magnitude, the magnitude of the centripetal acceleration simply depends on speed of the body.)

4.4.2 Circular Motion: Centripetal and Centrifugal Forces

From the value of the centripetal acceleration—the acceleration of a body moving at constant speed in a circle—we can use Newton's second law to find the value of the "centripetal force"—the net force needed to keep an object moving in a circle. Like the centripetal acceleration, this force is always directed towards the center of the circle. Its magnitude is

$$F_{\text{centripetal}} = ma_{\text{centripetal}} = \frac{mv^2}{r}$$

where m is the mass of the object moving in a circle, v is its velocity at any point in time (the direction doesn't matter, since the square of a vector is the square of its magnitude), and r is the radius of the circle.

It is important to recognize that *the centripetal force is not a new force produced by motion in a circle*. Rather, it's the value of the net force on an object that results in motion in a circle. That is, an object moves in a circle at constant speed if and only if $\Sigma F = F_{\text{centripetal}}$. This net force can be provided in many different ways, such as by tension in a rope, the internal forces of rigid bodies (as in the merry-go-round), or gravity (as in orbits).

Example: Centripetal Force from Friction

A child who weighs 20. kg is standing on a merry-go-round at a radius of 4.0 m from the center, where their speed is 5.0 m/s. What coefficient of static friction is needed to provide the centripetal force to keep them in place?

The centripetal force needed to maintain circular motion is

$$F_{\text{centripetal}} = \frac{mv^2}{r} = \frac{(20.\text{ kg})(5.0\text{ m/s})^2}{4.0\text{ m}} = 125\text{ N}$$

The maximum strength of the friction force is the product of the coefficient of static friction and the normal force. On a horizontal surface, the normal force will simply be equal to the weight force, so the child's maximum frictional force will be

$$F_{\text{friction}} = mgC_{\text{friction}} = (20.\text{ kg})(9.8\text{ N/kg})(C_{\text{friction}}) = (196\text{ N})(C_{\text{friction}})$$

This is the maximum centripetal force that can be provided by friction. Thus, the child will be able to stand on the merry-go-round as long as the coefficient of static friction between their feet and the merry-go-round satisfies

$$(196\text{ N})(C_{\text{friction}}) \geq 125\text{ N}$$

ie, $C_{\text{friction}} \geq 0.64$. If the coefficient of friction is less than this, they will have to hold on to something.

Of course, from the point of view of the child in the previous example, the frictional force isn't pulling them in a circle. Instead, in their rotating (and thus accelerating, i.e. non-inertial) frame of reference, there is a force pushing them away from the center of the merry-go-round, and friction is merely holding them in place. This perceived force is called the centrifugal force.[18]

The centrifugal force is an example of what is called a "fictional force," similar to the "Coriolis force" that is responsible for the rotation of hurricanes. Fictional forces are artifacts of attempting to use Newton's laws of motion in reference frames that are accelerating: when viewed in a non-accelerating frame of reference—for example, a parent standing on the ground and watching their child on the merry-go-round—these forces vanish, and the motion attributed to them becomes simply a natural consequence of inertia: the child is not being pulled outward by a mysterious force, they're just naturally moving in a straight line because of inertia and need friction to pull them off this inertial path into a circle.

4.4.3 Newton's Law of Universal Gravitation

Many forms of circular motion are due to mechanical forces, such as the friction keeping the child on the merry-go-round in the previous section. However, gravity is also capable of providing centripetal forces, resulting in orbits.

In order to discuss orbits, though, we will need a more general definition of the forces produced by gravity. So far, we have been defining the weight force as a force pointing downward with a magnitude of mg, where m is the mass of the object the force is acting on and g is a constant. This definition is quite useful when working near the vicinity of Earth's surface, but it breaks down as one's distance from the Earth increases, or in the vicinity of other planets.

In fact, the equation $F = mg$ is a special case of a more general law, **Newton's Law of Universal Gravitation**, which states that the force experienced by an object with mass in the vicinity of another object with mass is directed towards the second object with a magnitude

$$F = G\frac{m_1 m_2}{r^2}$$

where $G = 6.673 \times 10^{-11}\ \mathrm{kg}^{-1}\mathrm{m}^3\mathrm{s}^{-2}$ is a fundamental constant called the "constant of universal gravitation" or "Newton's constant," m_1 and m_2 are the masses of the two objects, and r is the distance between the two objects.

At first glance, the law of universal gravitation seems to only be useful for describing the force between point-sized masses that don't have a measurable size. After all, if we want to apply this to the gravitational interaction between ourselves and the planet Earth, it is

[18] The confusing similarity between the words "centripetal" and "centrifugal" is an artifact of the days when all physicists knew Latin. Centripetal means "center-seeking" and centrifugal means "center-fleeing."

unclear what the distance between a person and Earth is: the distance between them and the ground they're standing on, the distance between them and the center of the Earth, or what?

While we should technically apply the law of gravitation to each atom in ourselves and the Earth separately and add the resulting forces, this is not really practical. Fortunately, it has been proven that the gravitational force between a spherical object and any object outside the surface of the sphere is the same as if all the mass of the spherical object were concentrated at its center. In other words, we can treat the Earth,, the Moon, the Sun, and most other astronomical bodies as point masses. Furthermore, we can treat an object as a point mass so long as its size is much smaller than the distance between the two objects, which means that we can treat human-sized objects on Earth's surface, or spacecraft, as point masses as well, since they are much smaller than the radii of the Earth and other astronomical bodies.

Example: Gravity at Earth's Surface

What is the weight force on an object of mass m located on Earth's surface? What if the object is 415 km above Earth's surface, the altitude at which the International Space Station orbits?

The weight force on an object of mass m on Earth's surface, according to the law of universal gravitation, is

$$F = G\frac{mM_{\text{Earth}}}{r_{\text{Earth}}^2} = (6.673\times10^{-11}\ \text{kg}^{-1}\text{m}^3\text{s}^{-2})\frac{m(5.972\times10^{24}\ \text{kg})}{(6.371\times10^6\ \text{m})^2} = (9.8\ \text{N/kg})m$$

This value is the same as the value we would predict by using $g = 9.8\ \text{m/s}^2$ as gravitational acceleration. In other words, Newton's law of universal gravitation predicts that objects near Earth's surface will accelerate at the same rate regardless of their mass.

415 km above Earth's surface, this becomes

$$F = G\frac{mM_{\text{Earth}}}{(r_{\text{Earth}}+415\ \text{km})^2} = (6.673\times10^{-11}\ \text{kg}^{-1}\text{m}^3\text{s}^{-2})\frac{m(5.972\times10^{24}\ \text{kg})}{(6.786\times10^6\ \text{m})^2} = (8.7\ \text{N/kg})m$$

As we can see, the weight force and acceleration due to gravity are noticeably weaker at the altitude of the International Space Station. However, they are still much too high to make one feel "weightless."

4.4.4 Orbits

The example in the previous section demonstrated that the law of universal gravitation simplifies to produce the constant free-fall acceleration we observe at Earth's surface. However, it also tells us that astronauts on the International Space Station should still have substantial gravitational accelerations. Why do they feel weightless?

As we've already observed, we don't directly feel the effect of gravity as our weight: we instead feel the force from whatever is supporting us and keeping us from being in free fall as our weight. Objects in orbit, such as space stations, are constantly in free fall, with no forces acting on them except gravity. Thus, their passengers feel weightless.

How can an object remain in free fall indefinitely without eventually falling to Earth's surface? The answer is that it can be launched on a trajectory on which Earth's gravity provides a free-fall acceleration equal to the centripetal acceleration needed to keep the object on a circular trajectory, which in turn keeps the altitude, and thus the free-fall acceleration, constant.

Example: Finding an Orbital Velocity

What does the velocity of the International Space Station need to be for it to remain in a stable orbit?

We know from the previous example that the force of gravity on an object at the International Space Station's altitude is $m(8.7 \text{ N/kg})$, where m is the mass of the object. The centripetal force needed to keep an object on a circular path is $m\frac{v^2}{r}$, where m is the object's mass, r is the radius of the circle, and v is the velocity of the object at any point in time.

Thus, the International Space Station will remain in orbit if $\frac{v^2}{r} = 8.7 \text{ N/kg}$. Since r here is 6.786×10^6 m, the Earth's radius plus the space station's altitude above Earth's surface, $v = 7.7$ km/s perpendicular to the Earth's radius.

At this speed, it will remain in free fall permanently, with the only force acting on it the gravity that provides the centripetal acceleration. Since people can't directly feel the gravitational force acting on them, astronauts on board do not feel the effects of the centripetal force, and thus do not observe a fictional centrifugal force pulling them away from Earth's surface, either.

It is possible to show, with somewhat more complicated math, that the law of universal gravitation also allows elliptical, parabolic, and hyperbolic orbits, and thus governs the motion of the planets and other astronomical bodies according to the same mathematical relationship as applies to falling objects near Earth's surface.

While this seems like a commonplace, expected fact today, at the time it was a major conceptual leap in physics. Previously, it had been believed that the laws of nature that applied on Earth's surface were completely different from the laws of nature that governed the motion of "perfect" astronomical bodies. Even Kepler, who first correctly determined the shapes of planetary orbits and their mathematical characteristics was convinced that their properties were in some way derived from the geometry of the five Platonic solids, rather than from the same sort of gravitational acceleration that caused a falling apple to hit the ground.

This leap of insight—that the same laws of nature apply everywhere in the universe—was a major milestone in the development of modern science, and is one of several reasons that Newton is often considered the greatest physicist of all time.

5 Momentum and Kinetic Energy

5.1 Kinetic Energy

5.2 Momentum

5.3 Elastic Collisions

5.4 Completely Inelastic Collisions