NO BULLSHIT

guide to

MATH & PHYSICS

by Ivan Savov

Minireference Publishing. Fourth edition (v4.0), second printing.
ISBN: 978-0-9920010-0-1
The author can be reached at `ivan.savov@gmail.com`.
hg changeset: 237:1462c9c5beb5.

Contents

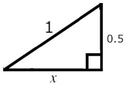
Placement exam

The answers[1] to this placement exam will tell you where to start reading.

1. What is the derivative of $\sin(x)$?

2. What is the second derivative of $A\sin(\omega x)$?

3. What is the value of x ?



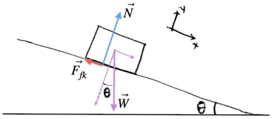
4. What is the magnitude of the gravitational force between two planets of mass M and mass m separated by a distance r?

5. Calculate $\lim_{x \to 3^-} \dfrac{1}{x-3}$.

6. Solve for t in:

$$7(3+4t) = 11(6t-4).$$

7. What is the component of the weight $\vec{W}$ acting in the x direction?



8. A mass-spring system is undergoing simple harmonic motion. Its position function is $x(t) = A\sin(\omega t)$. What is its maximum acceleration?

[1] Ans: 1. $\cos(x)$, 2. $-A\omega^2 \sin(\omega x)$, 3. $\frac{\sqrt{3}}{2}$, 4. $|\vec{F}_g| = \frac{GMm}{r^2}$, 5. $-\infty$, 6. $\frac{65}{38}$, 7. $+mg\sin\theta$, 8. $A\omega^2$. Key: If you didn't get Q3, Q6 right, you should read the book starting from Chapter 1. If you are mystified by Q1, Q2, Q5, read Chapter 5. If you want to learn how to solve Q4, Q7 and Q8, read Chapter 4.

Concept map

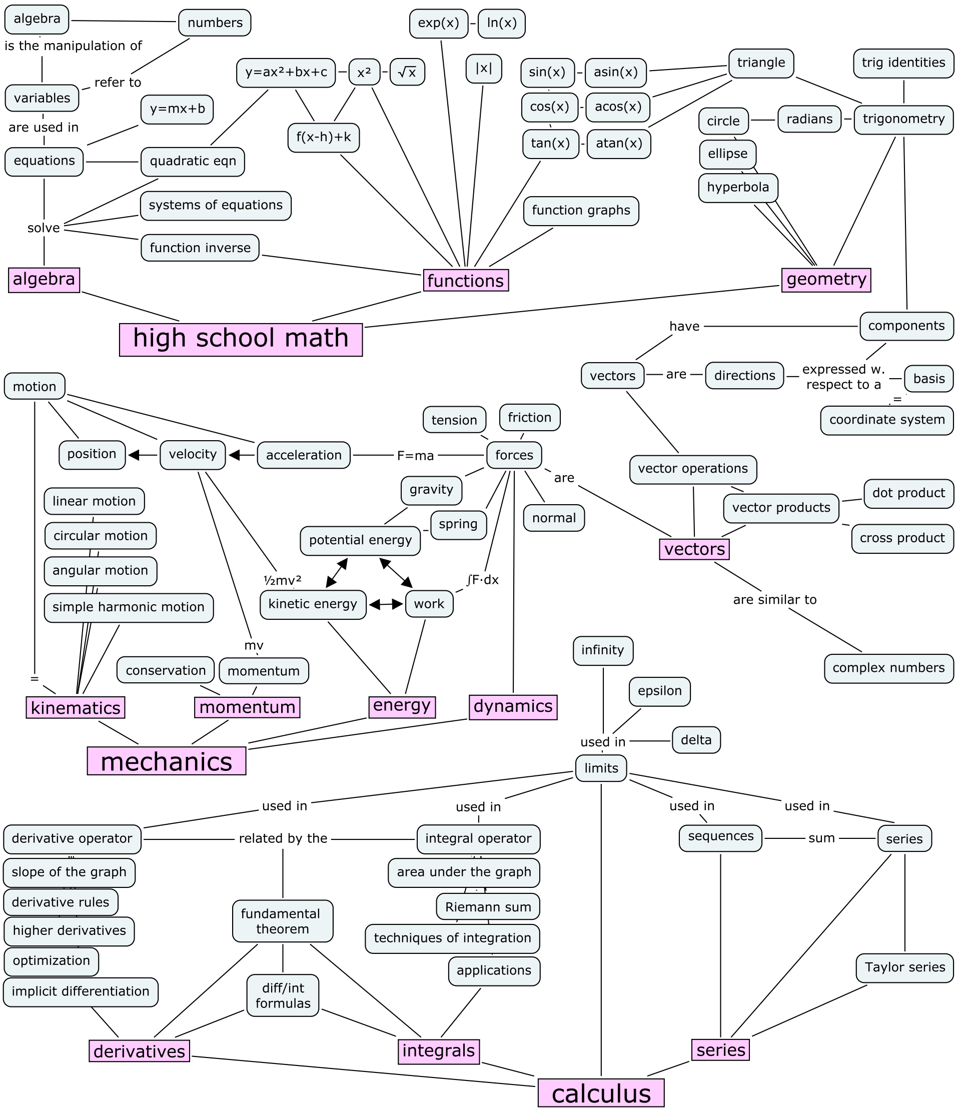


Figure 1: Each concept in this diagram corresponds to one section in the book.

Preface

This book contains lessons on topics in math and physics, written in a style that is jargon-free and to the point. Each lesson covers one concept at the depth required for a first-year university-level course. The main focus of this book is to highlight the intricate connections between the concepts of math and physics. Seeing the similarities and parallels between the concepts is the key to understanding.

Why?

The genesis of this book dates back to my student days when I was required to purchase expensive textbooks for my courses. Not only are these textbooks expensive, they are also tedious to read. Who has the energy to go through thousands of pages of explanations? I began to wonder, "What's the deal with these thick books?" Later, I realized mainstream textbooks are long because the textbook industry wants to make more profits. You don't need to read 1000 pages to learn calculus; the numerous full-page colour pictures and the repetitive text that are used to "pad" calculus textbooks are there to make the $\$130$ price seem reasonable.

Looking at this situation, I said to myself, "something must be done," and I sat down and wrote a modern textbook to explain math and physics clearly, concisely, and affordably. There was no way I was going to let mainstream publishers ruin the learning experience of these beautiful subjects for the next generation.

How?

Each section in this book is a **self-contained tutorial**. Each section covers the definitions, formulas, and explanations associated with a single topic. You can therefore read the sections in any order you find logical. Along the way, you will learn about the *connections* between the concepts of calculus and mechanics. Understanding mechanics is much easier if you know the ideas of calculus. At the same time, the ideas behind calculus are best illustrated through concrete physics examples. Learning the two subjects simultaneously is the best approach.

To learn mechanics and calculus, you first need to know your high school math. In order to make the book accessible for all readers, the book begins with a review chapter on numbers, algebra, equations, functions, and other prerequisite concepts. If you feel a little rusty on those concepts, be sure to check out Chapter 1.

The end of each section contains links to interesting webpages, animations, and further reading material. You can use these links as a starting point for further exploration. The end of each chapter contains a series of exercises. Make sure you spend some quality time with them. You will learn a lot by solving exercises on your own.

Is this book for you?

My aim is to make learning calculus and mechanics more accessible. Anyone should be able to open this book and become proficient in calculus and mechanics, regardless of their mathematical background.

The book's primary reader‘tudents. Students taking a mechanics class can read the chapters sequentially until Chapter 4, and optionally read Chapter 5 for general knowledge. Taking a calculus course? Skip ahead directly to the calculus chapter (Chapter 5). High school students or university students taking a precalculus class will benefit from reading Chapter 1, which is a concise but thorough review of fundamental math concepts like numbers, equations, functions, and trigonometry.

MECH CLASS	CALC CLASS	PRECALC CLASS
Ch. 1	Ch. 1	Ch. 1
Ch. 2	Ch. 2	Ch. 2†
Ch. 3		
Ch. 4		
Ch. 5†	Ch. 5	

† = optional reading.

Non-students, don't worry: you do not need to be taking a class in order to learn math. Independent learners interested in learning university-level material will find this book very useful. Many university graduates read this book to remember the calculus they learned back in their university days.

In general, anyone interested in rekindling and improving their relationship with mathematics should consider this book as an opportunity to repair the broken connection. Math is good stuff; you shouldn't miss out on it. People who think they absolutely *hate* math should read Chapter 1 as therapy.

About the author

I have been teaching math and physics for more than 10 years as a private tutor. My tutoring experience has taught me how to explain concepts that people find difficult to understand. I've had the chance to experiment with different approaches for explaining challenging material. Fundamentally, I've learned from teaching that understanding connections between concepts is much more important than memorizing facts. It's not about how many equations you know, but about knowing how to get from one equation to another.

I completed my undergraduate studies at McGill University in Electrical Engineering, then did a M.Sc. in Physics, and recently completed a Ph.D. in Computer Science. In my career as a researcher, I've been fortunate to learn from very inspirational teachers, who had the ability to distill the essential ideas and explain things in simple language. With my writing, I want to recreate the same learning experience for you. I founded the `Minireference Co.` to revolutionize the textbook industry. We make textbooks that don't suck.

Ivan Savov
Montreal, 2013

Introduction

The last two centuries have been marked by tremendous technological advances. Every sector of the economy has been transformed by the use of computers and the advent of the Internet. There is no doubt technology's importance will continue to grow in the coming years.

The best part is that you don't need to know how technology works to use it. You need not understand how Internet protocols operate to check your email and find original pirate material. You don't need to be a programmer to tell a computer to automate repetitive tasks and increase your productivity. However, when it comes to building *new* things, understanding becomes important. One particularly useful skill is the ability to create mathematical models of real-world situations. The techniques of mechanics and calculus are powerful building blocks for understanding the world around us. This is why these courses are taught in the first year of university studies: they contain keys that unlock the rest of science.

Calculus and mechanics can be difficult subjects. Understanding the material isn't hard *per se*, but it takes patience and practice. Calculus and mechanics become much easier to absorb when you break down the material into manageable chunks. It is most important you learn the *connections* between concepts.

Before we start with the equations, it's worthwhile to preview the material covered in this book. After all, you should know what kind of trouble you're getting yourself into.

Chapter 1 is a comprehensive review of math fundamentals including algebra, equation solving, and functions. The exposition of each topic is brief to make for easy reading. This chapter is highly recommended for readers who haven't looked at math recently; if you need a refresher on math, Chapter 1 is for you. It is extremely important to firmly grasp the basics. What is $\sin(0)$? What is $\sin(\pi/4)$? What does the graph of $\sin(x)$ look like? Arts students interested in enriching their cultural insight with knowledge that is 2000+ years old can read this chapter as therapy to recover from any damaging educational experiences they may have encountered in high school.

In Chapter 2, we'll look at how techniques of high school math can be used to describe and model the world. We'll learn about the basic laws that govern the motion of objects in one dimension and the mathematical equations that describe the motion. By the end of this chapter, you'll be able to predict the flight time of a ball thrown in the air.

In Chapter 3, we will learn about vectors. Vectors describe directional quantities like forces and velocities. We need vectors to properly understand the laws of physics. Vectors are used in many areas of science and technology, so becoming comfortable with vector calculations will pay dividends when learning other subjects.

Chapter 4 is all about mechanics. We'll study the motion of objects, predict their future trajectories, and learn how to use abstract concepts like momentum and energy. Science students who "hate" physics can study this chapter to learn how to use the 20 main equations and laws of physics. You will see physics is actually quite simple.

Chapter 5 covers topics from differential calculus and integral calculus. We will study limits, derivatives, integrals, sequences, and series. You will find that 100 pages are enough to cover all the concepts in calculus, as well as illustrate them with examples and practice exercises.

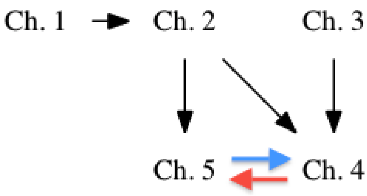


Figure 2: The prerequisite structure for the chapters in this book.

Calculus and mechanics are often taught as separate subjects. It shouldn't be like that! If you learn calculus without mechanics, it will be boring. If you learn physics without calculus, you won't truly understand. The exposition in this book covers both subjects in an integrated manner and aims to highlight the connections between them. Let's dig in.

Chapter 1

Math fundamentals

In this chapter we'll review the fundamental ideas of mathematics, including numbers, equations, and functions. To understand college-level textbooks, you need to be comfortable with mathematical calculations. Many people have trouble with math, however. Some people say they *hate* math, or could never learn it. It's not uncommon for children who score poorly on their school math exams to develop math complexes in their grown lives. If you are carrying any such emotional baggage, you can drop it right here and right now.

Do NOT worry about math! You are an adult, and you can learn math much more easily than when you were in high school. We'll review *everything* you need to know about high school math, and by the end of this chapter, you'll see that math is nothing to worry about.

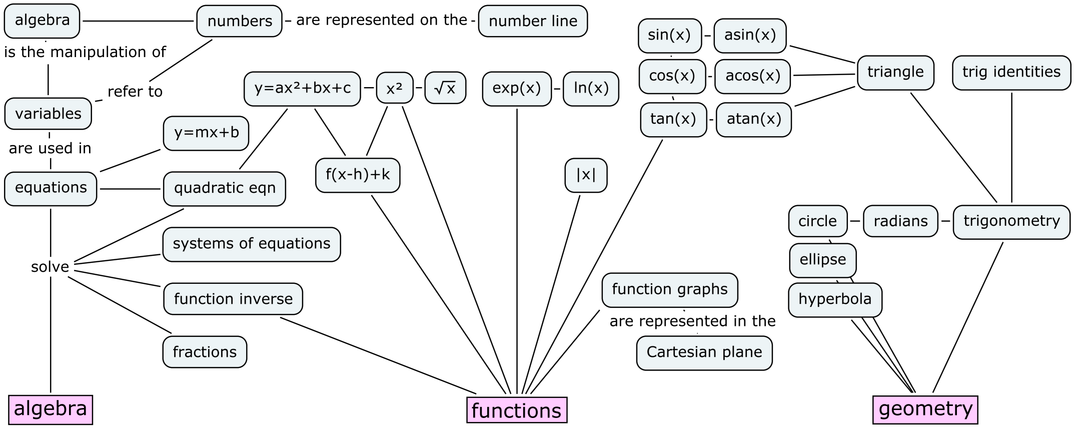


Figure 1.1: A concept map showing the mathematical topics that we will cover in this chapter. We'll learn about how to solve equations using algebra, how to model the world using functions, and how to think geometrically. The material in this chapter is required for your understanding of the more advanced topics in this book.

1.1 Solving equations

Most math skills boil down to being able to manipulate and solve equations. Solving an equation means finding the value of the unknown in the equation.

Check this shit out:

$$x^2 - 4 = 45.$$

To solve the above equation is to answer the question "What is x?" More precisely, we want to find the number that can take the place of x in the equation so that the equality holds. In other words, we're asking,

"Which number times itself minus four gives 45?"

That is quite a mouthful, don't you think? To remedy this verbosity, mathematicians often use specialized mathematical symbols. The problem is that these specialized symbols can be very confusing. Sometimes even the simplest math concepts are inaccessible if you don't know what the symbols mean.

What are your feelings about math, dear reader? Are you afraid of it? Do you have anxiety attacks because you think it will be too difficult for you? Chill! Relax, my brothers and sisters. There's nothing to it. Nobody can magically guess what the solution to an equation is immediately. To find the solution, you must break the problem down into simpler steps.

To find x, we can manipulate the original equation, transforming it into a different equation (as true as the first) that looks like this:

$$x = \text{only numbers.}$$

That's what it means to *solve*. The equation is solved because you can type the numbers on the right-hand side of the equation into a calculator and obtain the numerical value of x that you're seeking.

By the way, before we continue our discussion, let it be noted: the equality symbol ($=$) means that all that is to the left of $=$ is equal to all that is to the right of $=$. To keep this equality statement true, **for every change you apply to the left side of the equation, you must apply the same change to the right side of the equation**.

To find x, we need to correctly manipulate the original equation into its final form, simplifying it in each step. The only requirement is that the manipulations we make transform one true equation into another true equation. Looking at our earlier example, the first simplifying step is to add the number four to both sides of the equation:

$$x^2 - 4 + 4 = 45 + 4,$$

which simplifies to

$$x^2 = 49.$$

The expression looks simpler, yes? How did I know to perform this operation? I was trying to "undo" the effects of the operation -4. We undo an operation by applying its *inverse*. In the case where the operation is subtraction of some amount, the inverse operation is the addition of the same amount. We'll learn more about function inverses in Section 1.4.

We're getting closer to our goal, namely to *isolate* x on one side of the equation, leaving only numbers on the other side. The next step is to undo the square x^2 operation. The inverse operation of squaring a number x^2 is to take the square root $\sqrt{}$ so this is what we'll do next. We obtain

$$\sqrt{x^2} = \sqrt{49}.$$

Notice how we applied the square root to both sides of the equation? If we don't apply the same operation to both sides, we'll break the equality!

The equation $\sqrt{x^2} = \sqrt{49}$ simplifies to

$$|x| = 7.$$

What's up with the vertical bars around x? The notation $|x|$ stands for the *absolute value* of x, which is the same as x except we ignore the sign. For example $|5| = 5$ and $|-5| = 5$, too. The equation $|x| = 7$ indicates that both $x = 7$ and $x = -7$ satisfy the equation $x^2 = 49$. Seven squared is 49, and so is $(-7)^2 = 49$ because two negatives cancel each other out.

We're done since we isolated x. The final solutions are

$$x = 7 \qquad \text{or} \qquad x = -7.$$

Yes, there are *two* possible answers. You can check that both of the above values satisfy our initial equation $x^2 - 4 = 45$.

If you are comfortable with all the notions of high school math and you feel you could have solved the equation $x^2 - 4 = 25$ on your own, then you should consider skipping ahead to Chapter 2. If on the other hand you are wondering how the squiggle killed the power two, then this chapter is for you! In the next sections we will review all the essential concepts from high school math that you will need to power through the rest of this book. First, let me tell you about the different kinds of numbers.

1.2 Numbers

In the beginning, we must define the main players in the world of math: numbers.

Definitions

Numbers are the basic objects we use to calculate things. Mathematicians like to classify the different kinds of number-like objects into *sets*:

- The natural numbers: $\mathbb{N} = \{0, 1, 2, 3, 4, 5, 6, 7, \ldots\}$
- The integers: $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
- The rational numbers: $\mathbb{Q} = \{-1, 0, 0.125, 1, 1.5, \frac{5}{3}, \frac{22}{7}, \ldots\}$
- The real numbers: $\mathbb{R} = \{-1, 0, 1, e, \pi, -1.539\ldots,\ 4.94\ldots,\ \ldots\}$
- The complex numbers: $\mathbb{C} = \{-1, 0, 1, i, 1+i, 2+3i, \ldots\}$

These categories of numbers should be somewhat familiar to you. Think of them as neat classification labels for everything that you would normally call a number. Each item in the above list is a *set*. A set is a collection of items of the same kind. Each collection has a name and a precise definition. Note also that each of the sets in the list *contains* all the sets above it. For now, we don't need to go into the details of sets and set notation, but we do need to be aware of the different sets of numbers.

Why do we need so many different sets of numbers? The answer is partly historical and partly mathematical. Each set of numbers is associated with more and more advanced mathematical problems.

The simplest numbers are the natural numbers $\mathbb{N}$, which are sufficient for all your math needs if all you are going to do is *count* things. How many goats? Five goats here and six goats there so the total is 11 goats. The sum of any two natural numbers is also a natural number.

As soon as you start using *subtraction* (the inverse operation of addition), you start running into negative numbers, which are numbers outside the set of natural numbers. If the only mathematical operations you will ever use are *addition* and *subtraction*, then the set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ will be sufficient. Think about it. Any integer plus or minus any other integer is still an integer.

You can do a lot of interesting math with integers. There is an entire field in math called *number theory* that deals with integers. However, to restrict yourself solely to integers is somewhat limiting. You can't use the notion of 2.5 goats for example. The menu at Rotisserie Romados, which offers $\frac{1}{4}$ of a chicken, would be completely confusing.

If you want to use division in your mathematical calculations, you'll need the rationals $\mathbb{Q}$. The rationals are the set of *fractions* of integers:

$$\mathbb{Q} = \left\{ \text{all } z \text{ such that } z = \frac{x}{y} \text{ where } x \text{ and } y \text{ are in } \mathbb{Z}, \text{ and } y \neq 0 \right\}.$$

You can add, subtract, multiply, and divide rational numbers, and the result will always be a rational number. However, even the rationals are not enough for all of math!

In geometry, we can obtain *irrational* quantities like $\sqrt{2}$ (the diagonal of a square with side 1) and π (the ratio between a circle's circumference and its diameter). There are no integers x and y such that $\sqrt{2} = \frac{x}{y}$. Therefore, $\sqrt{2}$ is not part of the set $\mathbb{Q}$, and we say that $\sqrt{2}$ is *irrational*. An irrational number has an infinitely long decimal expansion that doesn't repeat. For example, $\pi = 3.1415926535897931\ldots$ where the dots indicate that the decimal expansion of π continues all the way to infinity.

Adding the irrational numbers to the rationals gives us all the useful numbers, which we call the set of real numbers $\mathbb{R}$. The set $\mathbb{R}$ contains the integers, the fractions $\mathbb{Q}$, as well as irrational numbers like $\sqrt{2} = 1.4142135\ldots$. By using the reals you can compute pretty much anything you want. From here on in the text, when I say *number*, I mean an element of the set of real numbers $\mathbb{R}$.

The only thing you can't do with the reals is take the square root of a negative number—you need the complex numbers $\mathbb{C}$ for that. We defer the discussion on $\mathbb{C}$ until the end of Chapter 3.

Operations on numbers

Addition

You can add and subtract numbers. I will assume you are familiar with this kind of stuff:

$$2+5=7, \quad 45+56=101, \quad 65-66=-1, \quad 9\,999+1=10\,000.$$

It can help visual learners to picture numbers as lengths measured out on the *number line*. Adding numbers is like adding sticks together: the resulting stick has a length equal to the sum of the lengths of the constituent sticks.

Addition is *commutative*, which means that $a+b=b+a$. It is also *associative*, which means that if you have a long summation like $a+b+c$ you can compute it in any order $(a+b)+c$ or $a+(b+c)$ and you'll get the same answer.

Subtraction is the inverse operation of addition.

Multiplication

You can also multiply numbers together.

$$ab = \underbrace{a+a+\cdots+a}_{b \ times} = \underbrace{b+b+\cdots+b}_{a \ times}.$$

Note that multiplication can be defined in terms of repeated addition.

The visual way to think about multiplication is as an area calculation. The area of a rectangle of base a and height b is equal to ab. A rectangle with a height equal to its base is a square, and this is why we call $aa=a^2$ "a squared."

Multiplication of numbers is also commutative, $ab=ba$; and associative, $abc=(ab)c=a(bc)$. In modern notation, no special symbol is used to denote multiplication; we simply put the two factors next to each other and say the multiplication is *implicit*. Some other ways to denote multiplication are $a \cdot b$, $a \times b$, and, on computer systems, $a * b$.

Division

Division is the inverse operation of multiplication.

$$a/b = \frac{a}{b} = \text{ one } b^{\text{th}} \text{ of } a.$$

Whatever a is, you need to divide it into b equal parts and take one such part. Some texts denote division as $a \div b$.

Note that you cannot divide by 0. Try it on your calculator or computer. It will say "`error divide by zero`" because this action simply doesn't make sense. After all, what would it mean to divide something into zero equal parts?

Exponentiation

Often an equation calls for us to multiply things together many times. The act of multiplying a number by itself many times is called *exponentiation*, and we denote this operation as a superscript:

$$a^b = \underbrace{aaa \cdots a}_{b \ times}.$$

We can also encounter negative exponents. The negative in the exponent does not mean "subtract," but rather "divide by":

$$a^{-b} = \frac{1}{a^b} = \frac{1}{\underbrace{aaa \cdots a}_{b \ times}}.$$

Fractional exponents describe square-root-like operations:

$$a^{\frac{1}{2}} \equiv \sqrt{a} \equiv \sqrt[2]{a}, \quad a^{\frac{1}{3}} \equiv \sqrt[3]{a}, \quad a^{\frac{1}{4}} \equiv \sqrt[4]{a} = a^{\frac{1}{2}\frac{1}{2}} = \left(a^{\frac{1}{2}}\right)^{\frac{1}{2}} = \sqrt{\sqrt{a}}.$$

Square root $\sqrt{x}$ is the inverse operation of x^2. Similarly, for any n we define the function $\sqrt[n]{x}$ (the n^{th} root of x) to be the inverse function of x^n.

It's worth clarifying what "taking the n^{th} root" means and understanding when to use this operation. The n^{th} root of a is a number which, when multiplied together n times, will give a. For example, a cube root satisfies

$$\sqrt[3]{a}\sqrt[3]{a}\sqrt[3]{a} = \left(\sqrt[3]{a}\right)^3 = a = \sqrt[3]{a^3}.$$

Do you see why $\sqrt[3]{x}$ and x^3 are inverse operations?

The fractional exponent notation makes the meaning of roots much more explicit. The n^{th} root of a can be denoted in two equivalent ways:

$$\sqrt[n]{a} \equiv a^{\frac{1}{n}}.$$

The symbol "$\equiv$" stands for "is equivalent to" and is used when two mathematical objects are identical. Equivalence is a stronger relation than equality. Writing $\sqrt[n]{a} = a^{\frac{1}{n}}$ indicates we've found two mathematical expressions (the left-hand side and the right-hand side of the equality) that happen to be equal to each other. It is more mathematically precise to write $\sqrt[n]{a} \equiv a^{\frac{1}{n}}$, which tells us $\sqrt[n]{a}$ and $a^{\frac{1}{n}}$ are two different ways of denoting the *same* mathematical object.

The n^{th} root of a is equal to one n^{th} of a with respect to multiplication. To find the whole number, multiply the number $a^{\frac{1}{n}}$ times itself n times:

$$\underbrace{a^{\frac{1}{n}}a^{\frac{1}{n}}a^{\frac{1}{n}}a^{\frac{1}{n}}\cdots a^{\frac{1}{n}}a^{\frac{1}{n}}}_{n \text{ times}} = \left(a^{\frac{1}{n}}\right)^n = a^{\frac{n}{n}} = a^1 = a.$$

The n-fold product of $\frac{1}{n}$-fractional exponents of any number produces that number with exponent one, therefore the inverse operation of $\sqrt[n]{x}$ is x^n.

The commutative law of multiplication $ab = ba$ implies that we can see any fraction $\frac{a}{b}$ in two different ways: $\frac{a}{b} = a\frac{1}{b} = \frac{1}{b}a$. We multiply by a then divide the result by b, or first we divide by b and then multiply the result by a. Similarly, when we have a fraction in the exponent, we can write the answer in two equivalent ways:

$$a^{\frac{2}{3}} = \sqrt[3]{a^2} = (\sqrt[3]{a})^2, \qquad a^{-\frac{1}{2}} = \frac{1}{a^{\frac{1}{2}}} = \frac{1}{\sqrt{a}}, \qquad a^{\frac{m}{n}} = \left(\sqrt[n]{a}\right)^m = \sqrt[n]{a^m}.$$

Make sure the above notation makes sense to you. As an exercise, try computing $5^{\frac{4}{3}}$ on your calculator and check that you obtain $8.54987973\ldots$ as the answer.

Operator precedence

There is a standard convention for the order in which mathematical operations must be performed. The basic algebra operations have the following precedence:

1. Exponents and roots
2. Products and divisions
3. Additions and subtractions

For instance, the expression $5 \times 3^2 + 13$ is interpreted as "first find the square of 3, then multiply it by 5, and then add 13." Parenthesis are needed to carry out the operations in a different order: to multiply 5 times 3 first and *then* take the square, the equation should read $(5 \times 3)^2 + 13$, where parenthesis indicate that the square acts on (5×3) as a whole and not on 3 alone.

Other operations

We can define all kinds of operations on numbers. The above three are special operations since they feel simple and intuitive to apply, but we can also define arbitrary transformations on numbers. We call these transformations *functions*. Before we learn about functions, let's first cover variables.

1.3 Variables

In math we use a lot of *variables*, which are placeholder names for *any* number or unknown.

Example Your friend invites you to a party and offers you to drink from a weirdly shaped shooter glass. You can't quite tell if it holds 25 ml of vodka or 50 ml or some amount in between. Since it's a mystery how much booze each shot contains, you shrug your shoulders and say there's x ml in there. The night happens. So how much did you drink? If you had three shots, then you drank $3x$ ml of vodka. If you want to take it a step further, you can say you drank n shots, making the total amount of alcohol you consumed nx ml.

Variables allow us to talk about quantities without knowing the details. This is *abstraction* and it is very powerful stuff: it allows you to get drunk without knowing how drunk exactly!

Variable names

There are common naming patterns for variables:

- x: general name for the unknown in equations (also used to denote a function's input, as well as an object's position in physics problems)
- v: velocity in physics problems
- θ, φ: the Greek letters *theta* and *phi* are often used to denote angles
- x_i, x_f: denotes an object's initial and final position in physics problems
- X: a random variable in probability theory
- C: costs in business along with P for profit, and R for revenue

Variable substitution

We can often *change variables* and replace one unknown variable with another to simplify an equation. For example, say you don't feel comfortable around square roots. Every time you see a square root, you freak out until one day you find yourself taking an exam trying to solve for x in the following equation:

$$\frac{6}{5-\sqrt{x}} = \sqrt{x}.$$

Don't freak out! In crucial moments like this, substitution can help with your root phobia. Just write, "Let $u = \sqrt{x}$" on your exam, and voila, you're allowed to rewrite the equation in terms of the variable u:

$$\frac{6}{5-u} = u,$$

which contains no square roots.

The next step to solve for u is to undo the division operation. Multiply both sides of the equation by $(5-u)$ to obtain

$$\frac{6}{5-u}(5-u) = u(5-u),$$

which simplifies to

$$6 = 5u - u^2.$$

This can be rewritten as a quadratic equation, $u^2 - 5u + 6 = 0$. Next, we can *factor* the quadratic to obtain the equation $(u-2)(u-3) = 0$, for which $u_1 = 2$ and $u_2 = 3$ are the solutions. The last step is to convert our u-answers into x answers by using $u = \sqrt{x}$, which is equivalent to $x = u^2$. The final answers are $x_1 = 2^2 = 4$ and $x_2 = 3^2 = 9$. Try plugging these x values into the original square root equation to verify that they satisfy it.

Compact notation

Symbolic manipulation is a powerful tool because it allows us to manage complexity. Say you're solving a physics problem in which you're told the mass of an object is $m = 140$ kg. If there are many steps in the calculation, would you rather use the number 140 kg in each step, or the shorter variable m? It's much easier in the long run to use the variable m throughout your calculation, and wait until the last step to substitute the value 140 kg when computing the final answer.

1.4 Functions and their inverses

As we saw in the section on solving equations, the ability to “undo” functions is a key skill for solving equations.

Example Suppose we’re solving for x in the equation

$$f(x) = c,$$

where f is some function and c is some constant. Our goal is to isolate x on one side of the equation, but the function f stands in our way.

By using the inverse function (denoted f^{-1}) we “undo” the effects of f. Then we apply the inverse function f^{-1} to both sides of the equation to obtain

$$f^{-1}(f(x)) = x = f^{-1}\left(c\right).$$

By definition, the inverse function f^{-1} performs the opposite action of the function f so together the two functions cancel each other out. We have $f^{-1}(f(x)) = x$ for any number x.

Provided everything is kosher (the function f^{-1} must be defined for the input c), the manipulation we made above is valid and we have obtained the answer $x = f^{-1}(c)$.

The above example introduces the notation f^{-1} for denoting the function’s *inverse*. This notation is borrowed from the notion of inverse numbers: multiplication by the number a^{-1} is the inverse operation of multiplication by the number a: $a^{-1}ax = 1x = x$. In the case of functions, however, the negative-one exponent does not refer to “one over-$f(x)$” as in $\frac{1}{f(x)} = (f(x))^{-1}$; rather, it refers to the function’s inverse. In other words, the number $f^{-1}(y)$ is equal to the number x such that $f(x) = y$.

Be careful: sometimes applying the inverse leads to multiple solutions. For example, the function $f(x) = x^2$ maps two input values (x and $-x$) to the same output value $x^2 = f(x) = f(-x)$. The inverse function of $f(x) = x^2$ is $f^{-1}(x) = \sqrt{x}$, and both $x = +\sqrt{c}$ and $x = -\sqrt{c}$ are solutions to the equation $x^2 = c$. In this case, this equation’s solutions can be indicated in shorthand notation as $x = \pm\sqrt{c}$.

Formulas

Here is a list of common functions and their inverses:

$$
\begin{array}{rcl}
\text{function } f(x) & \Leftrightarrow & \text{inverse } f^{-1}(x) \\
x+2 & \Leftrightarrow & x-2 \\
2x & \Leftrightarrow & \frac{1}{2}x \\
-x & \Leftrightarrow & -x \\
x^2 & \Leftrightarrow & \pm\sqrt{x} \\
2^x & \Leftrightarrow & \log_2(x) \\
3x+5 & \Leftrightarrow & \frac{1}{3}(x-5) \\
a^x & \Leftrightarrow & \log_a(x) \\
\exp(x) \equiv e^x & \Leftrightarrow & \ln(x) \equiv \log_e(x) \\
\sin(x) & \Leftrightarrow & \sin^{-1}(x) \equiv \arcsin(x) \\
\cos(x) & \Leftrightarrow & \cos^{-1}(x) \equiv \arccos(x)
\end{array}
$$

The function-inverse relationship is *reflexive*—if you see a function on one side of the above table (pick a side, any side), you'll find its inverse on the opposite side.

Example

Let's say your teacher doesn't like you and right away, on the first day of class, he gives you a serious equation and tells you to find x:

$$\log_5\left(3+\sqrt{6\sqrt{x}-7}\right) = 34 + \sin(5.5) - \Psi(1).$$

See what I mean when I say the teacher doesn't like you?

First, note that it doesn't matter what Ψ (the capital Greek letter *psi*) is, since x is on the other side of the equation. You can keep copying $\Psi(1)$ from line to line, until the end, when you throw the ball back to the teacher. "My answer is in terms of *your* variables, dude. *You* go figure out what the hell Ψ is since you brought it up in the first place!" By the way, it's not actually recommended to quote me verbatim should a situation like this arise. The same goes with $\sin(5.5)$. If you don't have a calculator handy, don't worry about it. Keep the expression $\sin(5.5)$ instead of trying to find its numerical value. In general, try to work with variables as much as possible and leave the numerical computations for the last step.

Okay, enough beating about the bush. Let's just find x and get it over with! On the right-hand side of the equation, we have the sum of a bunch of terms with no x in them, so we'll leave them as they are. On the left-hand side, the outermost function is a logarithm base 5. Cool. Looking at the table of inverse functions we find the exponential function is the inverse of the logarithm: $a^x \Leftrightarrow \log_a(x)$. To get rid of $\log_5$, we must apply the exponential function base 5 to both sides:

$$5^{\log_5\left(3+\sqrt{6\sqrt{x}-7}\right)} = 5^{34+\sin(5.5)-\Psi(1)},$$

which simplifies to

$$3 + \sqrt{6\sqrt{x}-7} = 5^{34+\sin(5.5)-\Psi(1)},$$

since 5^x cancels $\log_5 x$.

From here on, it is going to be as if Bruce Lee walked into a place with lots of bad guys. Addition of 3 is undone by subtracting 3 on both sides:

$$\sqrt{6\sqrt{x}-7} = 5^{34+\sin(5.5)-\Psi(1)} - 3.$$

To undo a square root we take the square:

$$6\sqrt{x} - 7 = \left(5^{34+\sin(5.5)-\Psi(1)} - 3\right)^2.$$

Add 7 to both sides,

$$6\sqrt{x} = \left(5^{34+\sin(5.5)-\Psi(1)} - 3\right)^2 + 7,$$

divide by 6

$$\sqrt{x} = \frac{1}{6}\left(\left(5^{34+\sin(5.5)-\Psi(1)} - 3\right)^2 + 7\right),$$

and square again to find the final answer:

$$x = \left[\frac{1}{6}\left(\left(5^{34+\sin(5.5)-\Psi(1)} - 3\right)^2 + 7\right)\right]^2.$$

Did you see what I was doing in each step? Next time a function stands in your way, hit it with its inverse so it knows not to challenge you ever again.

Discussion

The recipe I have outlined above is not universally applicable. Sometimes x isn't alone on one side. Sometimes x appears in several places in the same equation. In these cases, you can't effortlessly work your way, Bruce Lee-style, clearing bad guys and digging toward x—you need other techniques.

The bad news is there's no general formula for solving complicated equations. The good news is the above technique of "digging toward x" is sufficient for 80% of what you are going to be doing. You can get another 15% if you learn how to solve the quadratic equation:

$$ax^2 + bx + c = 0.$$

Solving third-degree polynomial equations like $ax^3 + bx^2 + cx + d = 0$ with pen and paper is also possible, but at this point you might as well start using a computer to solve for the unknowns.

There are all kinds of other equations you can learn how to solve: equations with multiple variables, equations with logarithms, equations with exponentials, and equations with trigonometric functions. The principle of "digging" toward the unknown by applying the function inverse is the key for solving all these types of equations, so be sure to practice it.

1.5 Basic rules of algebra

It's important that you know the general rules for manipulating numbers and variables, a process otherwise known as—you guessed it—*algebra*. This little refresher will cover these concepts to make sure you're comfortable on the algebra front. We'll also review some important algebraic tricks, like *factoring* and *completing the square*, which are useful when solving equations.

When an expression contains multiple things added together, we call those things *terms*. Furthermore, terms are usually composed of many things multiplied together. When a number x is obtained as the product of other numbers like $x = abc$, we say "x factors into a, b, and c." We call a, b, and c the *factors* of x.

Given any four numbers a, b, c, and d, we can apply the following algebraic properties:

1. Associative property: $a+b+c = (a+b)+c = a+(b+c)$ and $abc = (ab)c = a(bc)$
2. Commutative property: $a + b = b + a$ and $ab = ba$
3. Distributive property: $a(b + c) = ab + ac$

We use the distributive property every time we *expand* brackets. For example $a(b+c+d) = ab+ac+ad$. The brackets, also known as parentheses, indicate the expression $(b + c + d)$ must be treated as a whole: a factor that consists of three terms. Multiplying this expression by a is the same as multiplying each term by a.

The opposite operation of expanding is called *factoring*, which consists of rewriting the expression with the common parts taken out in front of a bracket: $ab + ac = a(b + c)$. In this section, we'll discuss both of these operations and illustrate what they're capable of.

Expanding brackets

The distributive property is useful when dealing with polynomials:

$$(x + 3)(x + 2) = x(x + 2) + 3(x + 2) = x^2 + x2 + 3x + 6.$$

We can use the commutative property on the second term $x2 = 2x$, then combine the two x terms into a single term to obtain

$$(x+3)(x+2) = x^2 + 5x + 6.$$

Let's look at this operation in its abstract form:

$$(x+a)(x+b) = x^2 + (a+b)x + ab.$$

The product of two linear terms (expressions of the form $x + ?$) is equal to a quadratic expression. Observe that the middle term on the right-hand side contains the *sum* of the two constants on the left-hand side $(a+b)$, while the third term contains their product ab.

It is very common for people to confuse these terms. If you are ever confused about an algebraic expression, go back to the distributive property and expand the expression using a step-by-step approach. As a second example, consider this slightly-more-complicated algebraic expression and its expansion:

$$\begin{aligned}(x+a)(bx^2+cx+d) &= x(bx^2+cx+d) + a(bx^2+cx+d)\\ &= bx^3 + cx^2 + dx + abx^2 + acx + ad\\ &= bx^3 + (c+ab)x^2 + (d+ac)x + ad.\end{aligned}$$

Note how all terms containing x^2 are grouped into a one term, and all terms containing x are grouped into another term. We use this pattern when dealing with expressions containing different powers of x.

Example Suppose we are asked to solve for t in the equation

$$7(3+4t) = 11(6t-4).$$

Since the unknown t appears on both sides of the equation, it is not immediately obvious how to proceed.

To solve for t, we must bring all t terms to one side and all constant terms to the other side. First, expand the two brackets to obtain

$$21 + 28t = 66t - 44.$$

Then move things around to relocate all ts to the equation's right-hand side and all constants to the left-hand side:

$$21 + 44 = 66t - 28t.$$

We see t is contained in both terms on the right-hand side, so we can rewrite the equation as

$$21 + 44 = (66 - 28)t.$$

The answer is within close reach: $t = \frac{21+44}{66-28} = \frac{65}{38}$.

Factoring

Factoring involves taking out the common part(s) of a complicated expression in order to make the expression more compact. Suppose you're given the expression $6x^2y + 15x$ and must simplify it by taking out common factors. The expression has two terms and each term can be split into its constituent factors to obtain

$$6x^2y + 15x = (3)(2)(x)(x)y + (5)(3)x.$$

Since factors x and 3 appear in both terms, we can *factor them out* to the front like this:

$$6x^2y + 15x = 3x(2xy + 5).$$

The expression on the right shows $3x$ is common to both terms.

Here's another example where factoring is used:

$$2x^2y + 2x + 4x = 2x(xy + 1 + 2) = 2x(xy + 3).$$

Quadratic factoring

When dealing with a quadratic function, it is often useful to rewrite the function as a product of two factors. Suppose you're given the quadratic function $f(x) = x^2 - 5x + 6$ and asked to describe its properties. What are the *roots* of this function? In other words, for what values of x is this function equal to zero? For which values of x is the function positive, and for which x values is the function negative?

Factoring the expression $x^2 + 5x + 6$ will help us see the properties of the function more clearly. To *factor* a quadratic expression is to express it as product of two factors:

$$f(x) = x^2 - 5x + 6 = (x - 2)(x - 3).$$

We now see at a glance the solutions (roots) are $x_1 = 2$ and $x_2 = 3$. We can also see for which x values the function will be overall positive: for $x > 3$, both factors will be positive, and for $x < 2$ both factors will be negative, and a negative times a negative gives a positive. For values of x such that $2 < x < 3$, the first factor will be positive, and the second factor negative, making the overall function negative.

For certain simple quadratics like the one above, you can simply *guess* what the factors will be. For more complicated quadratic expressions, you'll need to use the quadratic formula, which will be the subject of the next section. For now let us continue with more algebra tricks.

Completing the square

Any quadratic expression Ax^2+Bx+C can be rewritten in the form $A(x-h)^2+k$ for some constants h and k. This process is called *completing the square* due to the reasoning we follow to find the value of k. The constants h and k can be interpreted geometrically as the horizontal and vertical shifts in the graph of the basic quadratic function. The graph of the function $f(x) = A(x - h)^2 + k$ is the same as the graph of the function $f(x) = Ax^2$ except it is shifted h units to the right and k units upward. We will discuss the geometrical meaning of h and k in more detail in Section 1.14 (page 72). For now, let's focus on the algebra steps.

Let's try to find the values of k and h needed to complete the square in the expression $x^2 + 5x + 6$. We start from the assumption that the two expressions are equal, and then expand the bracket to obtain

$$\underline{x^2} + 5x + 6 = A(x - h)^2 + k = A(x^2 - 2hx + h^2) + k = \underline{Ax^2} - 2Ahx + Ah^2 + k.$$

Observe the structure in the above equation. On both sides of the equality there is one term which contains x^2 (the quadratic term), one term that contains x^1 (the linear term), and some constant terms. By focusing on the quadratic terms on both sides of the equation (they are underlined) we see $A = 1$, so we can rewrite the equation as

$$x^2 + \underline{5x} + 6 = x^2\underline{-2hx} + h^2 + k.$$

Next we look at the linear terms (underlined) and infer $h = -2.5$. After rewriting, we obtain an equation with a single unknown:

$$x^2 + 5x + \underline{6} = x^2 - 2(-2.5)x + \underline{(-2.5)^2 + k}.$$

Finally, we pick a value of k that will make the constant terms match:

$$k = 6 - (-2.5)^2 = 6 - (2.5)^2 = 6 - \left(\frac{5}{2}\right)^2 = 6 \times \frac{4}{4} - \frac{25}{4} = \frac{24 - 25}{4} = \frac{-1}{4}.$$

After completing the square we obtain

$$x^2 + 5x + 6 = (x + 2.5)^2 - \frac{1}{4}.$$

The right-hand side of the expression above tells us our function is equivalent to the basic function x^2, shifted 2.5 units to the left and $\frac{1}{4}$ units down. This would be very useful information if you ever had to draw the graph of this function—you could simply plot the basic graph of x^2 and then shift it appropriately.

It is important you become comfortable with this procedure for completing the square. It is not extra difficult, but it does require you to think carefully about the unknowns h and k and to choose their values appropriately. There is no general formula for finding k, but you can remember the following simple shortcut for finding h. Given an equation $Ax^2 + Bx + C = A(x - h)^2 + k$, we have $h = \frac{-B}{2A}$. Using this shortcut will save you some time, but you will still have to go through the algebra steps to find k.

Take out a pen and a piece of paper now (yes, right now!) and verify that you can correctly complete the square in these expressions: $x^2 - 6x + 13 = (x - 3)^2 + 4$ and $x^2 + 4x + 1 = (x + 2)^2 - 3$.

1.6 Solving quadratic equations

What would you do if asked to solve for x in the quadratic equation $x^2 = 45x + 23$? This is called a *quadratic equation* since it contains the unknown variable x squared. The name comes from the Latin *quadratus*, which means square. Quadratic equations appear often, so mathematicians created a general formula for solving them. In this section, we'll learn about this formula and use it to put some quadratic equations in their place.

Before we can apply the formula, we need to rewrite the equation we are trying to solve in the following form:

$$ax^2 + bx + c = 0.$$

We reach this form—called the *standard form* of the quadratic equation—by moving all the numbers and xs to one side and leaving only 0 on the other side. For example, to transform the quadratic expression $x^2 = 45x + 23$ into standard form, subtract $45x + 23$ from both sides of the equation to obtain $x^2 - 45x - 23 = 0$. What are the values of x that satisfy this formula?

Claim

The solutions to the equation $ax^2 + bx + c = 0$ are

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad \text{and} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Let's see how these formulas are used to solve $x^2 - 45x - 23 = 0$. Finding the two solutions requires the simple mechanical task of identifying $a = 1$, $b = -45$, and $c = -23$ and plugging these values into the formulas:

$$x_1 = \frac{45 + \sqrt{45^2 - 4(1)(-23)}}{2} = 45.5054\ldots,$$

$$x_2 = \frac{45 - \sqrt{45^2 - 4(1)(-23)}}{2} = -0.5054\ldots.$$

Verify using your calculator that both of the values above satisfy the original equation $x^2 = 45x + 23$.

Proof of claim

This is an important proof. I want you to see how we can *derive* the quadratic formula from first principles because this knowledge will help you understand the formula. The proof will use the completing-the-square technique from the previous section.

Starting with $ax^2 + bx + c = 0$, first move c to the other side of the equation:

$$ax^2 + bx = -c.$$

Divide by a on both sides:

$$x^2 + \frac{b}{a}x = -\frac{c}{a}.$$

Now *complete the square* on the left-hand side by asking, "What are the values of h and k that satisfy this equation

$$(x-h)^2 + k = x^2 + \frac{b}{a}x \quad ?\text{''}$$

To find the values for h and k, we'll expand the left-hand side to obtain $(x-h)^2 + k = x^2 - 2hx + h^2 + k$. We can now identify h by looking at the coefficients in front of x on both sides of the equation. We have $-2h = \frac{b}{a}$ and hence $h = -\frac{b}{2a}$.

Let's see what we have so far:

$$\left(x + \frac{b}{2a}\right)^2 = \left(x + \frac{b}{2a}\right)\left(x + \frac{b}{2a}\right) = x^2 + \frac{b}{2a}x + x\frac{b}{2a} + \frac{b^2}{4a^2} = x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}.$$

To determine k, we need to move that last term to the other side:

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} = x^2 + \frac{b}{a}x.$$

We can continue with the proof where we left off:

$$x^2 + \frac{b}{a}x = -\frac{c}{a}.$$

Replace the left-hand side with the complete-the-square expression and obtain

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} = -\frac{c}{a}.$$

From here on, we can use the standard procedure for solving equations. Arrange all constants on the right-hand side:

$$\left(x + \frac{b}{2a}\right)^2 = -\frac{c}{a} + \frac{b^2}{4a^2}.$$

Next, take the square root of both sides. Since the square function maps both positive and negative numbers to the same value, this step yields two solutions:

$$x + \frac{b}{2a} = \pm\sqrt{-\frac{c}{a} + \frac{b^2}{4a^2}}.$$

Let's take a moment to tidy up the mess under the square root:

$$\sqrt{-\frac{c}{a} + \frac{b^2}{4a^2}} = \sqrt{-\frac{(4a)c}{(4a)a} + \frac{b^2}{4a^2}} = \sqrt{\frac{-4ac + b^2}{4a^2}} = \frac{\sqrt{b^2 - 4ac}}{2a}.$$

We obtain

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

which is just one step from the final answer,

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

This completes the proof. □

Alternative proof of claim

To have a proof we don't necessarily need to show the derivation of the formula as outlined above. The claim states that x_1 and x_2 are solutions. To prove the claim we can simply plug x_1 and x_2 into the quadratic equation and verify the answers are zero. **Verify this on your own.**

Applications

The Golden Ratio

The golden ratio is an essential proportion in geometry, art, aesthetics, biology, and mysticism, and is usually denoted as $\varphi = \frac{1+\sqrt{5}}{2} = 1.6180339\ldots$. This ratio is determined as the positive solution to the quadratic equation

$$x^2 - x - 1 = 0.$$

Applying the quadratic formula to this equation yields two solutions,

$$x_1 = \frac{1+\sqrt{5}}{2} = \varphi \qquad \text{and} \qquad x_2 = \frac{1-\sqrt{5}}{2} = -\frac{1}{\varphi}\,.$$

You can learn more about the various contexts in which the golden ratio appears from the Wikipedia article on the subject. We'll discuss the golden ratio again on page 392 in Chapter 5.

Explanations

Multiple solutions

Often, we are interested in only one of the two solutions to the quadratic equation. It will usually be obvious from the context of the problem which of the two solutions should be kept and which should be discarded. For example, the *time of flight* of a ball thrown in the air from a height of 3 metres with an initial velocity of 12 metres per second is obtained by solving the equation $(-4.9)t^2 + 12t + 3 = 0$. The two solutions of the quadratic equation are $t_1 = -0.229$ and $t_2 = 2.678$. The first answer t_1 corresponds to a time in the past so we reject it as invalid. The correct answer is t_2. The ball will hit the ground after $t = 2.678$ seconds.

Relation to factoring

In the previous section we discussed the *quadratic factoring* operation by which we could rewrite a quadratic function as the product of two terms $f(x) = ax^2 + bx + c = (x - x_1)(x - x_2)$. The two numbers x_1 and x_2 are called the *roots* of the function: these points are where the function $f(x)$ touches the x-axis.

You now have the ability to factor any quadratic equation. Use the quadratic formula to find the two solutions, x_1 and x_2, then rewrite the expression as $(x - x_1)(x - x_2)$.

Some quadratic expressions cannot be factored, however. These "unfactorable" expressions correspond to quadratic functions whose graphs do not touch the x-axis. They have no solutions (no roots). There is a quick test you can use to check if a quadratic function $f(x) = ax^2 + bx + c$ has roots (touches or crosses the x-axis) or doesn't have roots (never touches the x-axis). If $b^2 - 4ac > 0$ then the function f has two roots. If $b^2 - 4ac = 0$, the function has only one root, indicating the special case when the function touches the x-axis at only one point. If $b^2 - 4ac < 0$, the function has no roots. In this case the quadratic formula fails because it requires taking the square root of a negative number, which is not allowed. Think about it—how could you square a number and obtain a negative number?

1.7 Exponents

In math we must often multiply together the same number many times, so we use the notation

$$b^n = \underbrace{bbb \cdots bb}_{n \text{ times}}$$

to denote some number b multiplied by itself n times. In this section we'll review the basic terminology associated with exponents and discuss their properties.

Definitions

The fundamental ideas of exponents are:

- b^n: the number b raised to the power n
  - ▷ b: the *base*
  - ▷ n: the *exponent* or *power* of b in the expression b^n

By definition, the zeroth power of any number is equal to one, expressed as $b^0 = 1$.

We'll also discuss *exponential functions* of the form $f : \mathbb{R} \to \mathbb{R}$. In particular, we define the following important exponential functions:

- b^x: the exponential function base b
- 10^x: the exponential function base 10
- $\exp(x) \equiv e^x$: the exponential function base e. The number e is called *Euler's number*.
- 2^x: the exponential function base 2. This function is very important in computer science.

The number $e = 2.7182818\ldots$ is a special base with many applications. We call e the *natural* base.

Another special base is 10 because we use the decimal system for our numbers. We can write very large numbers and very small numbers as powers of 10. For example, one thousand can be written as $1\,000 = 10^3$, one million is $1\,000\,000 = 10^6$, and one billion is $1\,000\,000\,000 = 10^9$.

Formulas

The following properties follow from the definition of exponentiation as repeated multiplication.

Property 1 Multiplying together two exponential expressions that have the same base is the same as adding the exponents:

$$b^m b^n = \underbrace{bbb\cdots bb}_{m \text{ times}}\underbrace{bbb\cdots bb}_{n \text{ times}} = \underbrace{bbbbbbb\cdots bb}_{m+n \text{ times}} = b^{m+n}.$$

Property 2 Division by a number can be expressed as an exponent of minus one:

$$b^{-1} \equiv \frac{1}{b}.$$

A negative exponent corresponds to a division:

$$b^{-n} = \frac{1}{b^n}.$$

Property 3 By combining Property 1 and Property 2 we obtain the following rule:

$$\frac{b^m}{b^n} = b^{m-n}.$$

In particular we have $b^n b^{-n} = b^{n-n} = b^0 = 1$. Multiplication by the number b^{-n} is the inverse operation of multiplication by the number b^n. The net effect of the combination of both operations is the same as multiplying by one, i.e., the identity operation.

Property 4 When an exponential expression is exponentiated, the inner exponent and the outer exponent multiply:

$$(b^m)^n = \underbrace{(\underbrace{bbb\cdots bb}_{m \text{ times}})(\underbrace{bbb\cdots bb}_{m \text{ times}})\cdots(\underbrace{bbb\cdots bb}_{m \text{ times}})}_{n \text{ times}} = b^{mn}.$$

Property 5.1

$$(ab)^n = \underbrace{(ab)(ab)(ab)\cdots(ab)(ab)}_{n \text{ times}} = \underbrace{aaa\cdots aa}_{n \text{ times}}\underbrace{bbb\cdots bb}_{n \text{ times}} = a^n b^n.$$

Property 5.2

$$\left(\frac{a}{b}\right)^n = \underbrace{\left(\frac{a}{b}\right)\left(\frac{a}{b}\right)\left(\frac{a}{b}\right)\cdots\left(\frac{a}{b}\right)\left(\frac{a}{b}\right)}_{n \text{ times}} = \frac{\overbrace{aaa\cdots aa}^{n \text{ times}}}{\underbrace{bbb\ \cdots\ bb}_{n \text{ times}}} = \frac{a^n}{b^n}.$$

Property 6 Raising a number to the power $\frac{1}{n}$ is equivalent to finding the n^{th} root of the number:

$$b^{\frac{1}{n}} \equiv \sqrt[n]{b}.$$

In particular, the square root corresponds to the exponent of one half: $\sqrt{b} = b^{\frac{1}{2}}$. The cube root (the inverse of x^3) corresponds to $\sqrt[3]{b} \equiv b^{\frac{1}{3}}$. We can verify the inverse relationship between $\sqrt[3]{x}$ and x^3 by using either Property 1: $(\sqrt[3]{x})^3 = (x^{\frac{1}{3}})(x^{\frac{1}{3}})(x^{\frac{1}{3}}) = x^{\frac{1}{3}+\frac{1}{3}+\frac{1}{3}} = x^1 = x$, or by using Property 4: $(\sqrt[3]{x})^3 = (x^{\frac{1}{3}})^3 = x^{\frac{3}{3}} = x^1 = x$.

Properties 5.1 and 5.2 also apply for fractional exponents:

$$\sqrt[n]{ab} = (ab)^{\frac{1}{n}} = a^{\frac{1}{n}} b^{\frac{1}{n}} = \sqrt[n]{a}\sqrt[n]{b}, \qquad \sqrt[n]{\left(\frac{a}{b}\right)} = \left(\frac{a}{b}\right)^{\frac{1}{n}} = \frac{a^{\frac{1}{n}}}{b^{\frac{1}{n}}} = \frac{\sqrt[n]{a}}{\sqrt[n]{b}}.$$

Discussion

Even and odd exponents

The function $f(x) = x^n$ behaves differently depending on whether the exponent n is even or odd. If n is odd we have

$$\left(\sqrt[n]{b}\right)^n = \sqrt[n]{b^n} = b.$$

However, if n is even, the function x^n destroys the sign of the number (see x^2, which maps both $-x$ and x to x^2). The successive application of exponentiation by n and the n^{th} root has the same effect as the absolute value function:

$$\sqrt[n]{b^n} = |b|.$$

Recall that the absolute value function $|x|$ discards the information about the sign of x.

The expression $(\sqrt[n]{b})^n$ cannot be computed whenever b is a negative number. The reason is that we can't evaluate $\sqrt[n]{b}$ for $b < 0$ in terms of real numbers, since there is no real number which, multiplied times itself an even number of times, gives a negative number.

Scientific notation

In science we often work with very large numbers like *the speed of light* ($c = 299\,792\,458$[m/s]), and very small numbers like *the permeability of free space* ($\mu_0 = 0.000001256637\ldots$[N/A^2]). It can be difficult to judge the magnitude of such numbers and to carry out calculations on them using the usual decimal notation.

Dealing with such numbers is much easier if we use *scientific notation*. For example, the speed of light can be written as $c = 2.99792458 \times 10^8$[m/s], and the permeability of free space is denoted as $\mu_0 = 1.256637 \times 10^{-6}$[N/A^2]. In both cases, we express the number as a decimal number between 1.0 and $9.9999\ldots$ followed by the number 10 raised to some power. The effect of multiplying by 10^8 is to move the decimal point eight steps to the right, making the number bigger. Multiplying by 10^{-6} has the opposite effect, moving the decimal to the left by six steps and making the number smaller. Scientific notation is useful because it allows us to clearly see the *size* of numbers: 1.23×10^6 is $1\,230\,000$ whereas 1.23×10^{-10} is $0.000\,000\,000\,123$. With scientific notation you don't have to count the zeros! Cool, yeah?

The number of decimal places we use when specifying a certain physical quantity is usually an indicator of the *precision* with which we are able to measure this quantity. Taking into account the precision of the measurements we make is an important aspect of all quantitative research. Since elaborating further would be a digression, we will not go into a full discussion about the topic of *significant digits* here. Feel free to check out the Wikipedia article on the subject if you want to know more.

On computer systems, *floating point numbers* are represented in scientific notation: they have a decimal part and an exponent. To separate the decimal part from the exponent when entering a floating point number into the computer, use the character e, which stands for "exponent." The base is assumed to be 10. For example, the speed of light is written as 2.99792458e8 and the permeability of free space is 1.256637e-6.

Links

[Further reading on exponentiation]
http://en.wikipedia.org/wiki/Exponentiation

[More details on scientific notation]
http://en.wikipedia.org/wiki/Scientific_notation

1.8 Logarithms

Some people think the word "logarithm" refers to some mythical, mathematical beast. Legend has it that logarithms are many-headed, breathe fire, and are extremely difficult to understand. Nonsense! Logarithms are simple. It will take you at most a couple of pages to get used to manipulating them, and that is a good thing because logarithms are used all over the place.

The strength of your sound system is measured in logarithmic units called decibels $[\mathrm{dB}]$. This is because your ears are sensitive only to exponential differences in sound intensity. Logarithms allow us to compare very large numbers and very small numbers on the same scale. If sound were measured in linear units instead of logarithmic units, your sound system's volume control would need to range from 1 to $1\,048\,576$. That would be weird, no? This is why we use the logarithmic scale for volume notches. Using a logarithmic scale, we can go from sound intensity level 1 to sound intensity level $1\,048\,576$ in 20 "progressive" steps. Assume each notch doubles the sound intensity, rather than increasing the intensity by a fixed amount. If the first notch corresponds to 2, the second notch is 4—still probably inaudible, turn it up! By the time you get to the sixth notch you're at $2^6 = 64$

sound intensity, which is the level of audible music. The tenth notch corresponds to sound intensity $2^{10} = 1024$ (medium-strength sound), and finally the twentieth notch reaches a max power of $2^{20} = 1\,048\,576$, at which point the neighbours come knocking to complain.

Definitions

You are hopefully familiar with these following concepts from the previous section:

- b^x: the exponential function base b
- $\exp(x) = e^x$: the exponential function base e, Euler's number
- 2^x: exponential function base 2
- $f(x)$: the notion of a function $f : \mathbb{R} \to \mathbb{R}$
- $f^{-1}(y)$: the inverse function of $f(x)$. It is defined in terms of $f(x)$ such that $f^{-1}(f(x)) = x$. In other words, if you apply f to some number and get the output y, and then you pass y through f^{-1}, the output will be x again. The inverse function f^{-1} undoes the effects of the function f.

In this section we will play with the following new concepts:

- $\log_b(x)$: the logarithm of x base b is the inverse function of b^x.
- $\ln(x)$: the "natural" logarithm base e. This is the inverse of e^x.
- $\log_2(x)$: the logarithm base 2 is the inverse of 2^x.

I say *play* because there is nothing much new to learn here: a logarithm is a clever way to talk about the size of a number; essentially, it tells us how many digits the number has.

Formulas

The main thing to realize is that $\log$s don't really exist on their own. They are defined as the inverses of their corresponding exponential functions. The following statements are equivalent:

$$\log_b(x) = m \qquad \Leftrightarrow \qquad b^m = x.$$

Logarithms with base e are written $\ln(x)$ for "logarithme naturel" because e is the "natural" base. Another special base is 10 because our numbers are based on the decimal system. The logarithm base 10 $\log_{10}(x)$ tells us roughly the size of the number x—how many digits the number has.

Example When someone working for the System (say someone with a high-paying job in the financial sector) boasts about his or her "six-figure" salary, they are really talking about the $\log$ of how much money they make. The "number of figures" N_S in their salary is calculated as 1 plus the logarithm base 10 of their salary S. The formula is

$$N_S = 1 + \log_{10}(S).$$

A salary of $S = 100\,000$ corresponds to $N_S = 1 + \log_{10}(100\,000) = 1 + 5 = 6$ figures. What is the smallest "seven-figure" salary? We must solve for S given $N_S = 7$ in the formula. We find $7 = 1 + \log_{10}(S)$, which means $6 = \log_{10}(S)$, and—using the inverse relationship between logarithm base 10 and exponentiation base 10—we discover $S = 10^6 = 1\,000\,000$. One million dollars per year! Yes, for this kind of money I see how someone might want to work for the System. But most system pawns never make it to the seven-figure level; I believe the average high-ranking salary is more in the $1 + \log_{10}(250\,000) = 1 + 5.397 = 6.397$ digits range. Wait, a lousy 0.397 extra digits is all it takes to convince some of the smartest people out there to sell their brains to the finance sector? What wankers! Who needs a six-digit salary anyway? Why not make $1 + \log_{10}(44\,000) = 5.64$ digits as a teacher and do something with your life that *actually* matters?

Properties

Moving on, let's discuss two important properties you'll need when dealing with logarithms. Pay attention because the arithmetic rules for logarithms are very different from the usual rules for numbers. Intuitively, you can think of logarithms as a convenient way to refer to the exponents of numbers. The following properties are the logarithmic analogues of the properties of exponents.

Property 1

The first property states that the sum of two logarithms is equal to the logarithm of the product of the *arguments*:

$$\log(x) + \log(y) = \log(xy).$$

From this property, we can derive two other useful ones:

$$\log(x^k) = k\log(x),$$

and

$$\log(x) - \log(y) = \log\left(\frac{x}{y}\right).$$

Proof: For all three equations above, we need to show that the expression on the left is equal to the expression on the right. We met logarithms a very short time ago, so we don't know each other too well yet. In fact, the only thing we know about $\log$s is the inverse relationship with the exponential function. The only way to prove this property is to use this relationship.

The following statement is true for any base b:

$$b^m b^n = b^{m+n}.$$

This follows from first principles. Recall that exponentiation is nothing more than repeated multiplication. If you count the total number of bs multiplied on the left side, you'll find a total of $m + n$ of them, which is what we have on the right.

If we define some new variables x and y such that $b^m = x$ and $b^n = y$, then we can rewrite the equation $b^m b^n = b^{m+n}$ as

$$xy = b^{m+n}.$$

Taking the logarithm of both sides gives us

$$\log_b(xy) = \log_b\big(b^{m+n}\big) = m + n = \log_b(x) + \log_b(y).$$

The last step above uses the definition of the $\log$ function again, which states that

$$b^m = x \quad\Leftrightarrow\quad m = \log_b(x) \qquad \text{and} \qquad b^n = y \quad\Leftrightarrow\quad n = \log_b(y).$$

Property 2

This property helps us change from one base to another.

We can express the logarithm in any base B in terms of a ratio of logarithms in another base b. The general formula is

$$\log_B(x) = \frac{\log_b(x)}{\log_b(B)}.$$

For example, the logarithm base 10 of a number S can be expressed as a logarithm base 2 or base e as follows:

$$\log_{10}(S) = \frac{\log_{10}(S)}{1} = \frac{\log_{10}(S)}{\log_{10}(10)} = \frac{\log_2(S)}{\log_2(10)} = \frac{\ln(S)}{\ln(10)}.$$

This property will help if you ever need to compute a logarithm in a base that is not available on your calculator. Suppose you are asked to compute $\log_7(S)$, but your calculator only has a $\boxed{\log_{10}}$ button. You can simulate $\log_7(S)$ by computing $\log_{10}(S)$ and dividing by $\log_{10}(7)$.

1.9 Fractions

The set of rational numbers $\mathbb{Q}$ is the set of numbers that can be written as a *fraction* of two integers:

$$\mathbb{Q} \equiv \left\{ \frac{m}{n} \;\middle|\; m \text{ and } n \text{ are in } \mathbb{Z} \text{ and } n \neq 0 \right\},$$

where $\mathbb{Z}$ denotes the set of integers $\mathbb{Z} \equiv \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. Fractions describe what happens when a *whole* is cut into n equal parts and we are given m of those parts.

We read $\frac{1}{4}$ either as *one over four* or *one quarter*, which is also equal to 0.25, but as you can see the notation $\frac{1}{4}$ is more compact and nicer. Why nicer? Check out these simple fractions:

$$\frac{1}{1} = 1.0$$

$$\begin{aligned}
\frac{1}{2} &= 0.5\\
\frac{1}{3} &= 0.33333\ldots = 0.\overline{3}\\
\frac{1}{4} &= 0.25\\
\frac{1}{5} &= 0.2\\
\frac{1}{6} &= 0.166666\ldots = 0.1\overline{6}\\
\frac{1}{7} &= 0.142857142857714285\ldots = 0.\overline{142857}
\end{aligned}$$

Note that a line above some numbers means the digits underneath the line are repeated. The fractional notation on the left is preferable, because it shows the underlying *structure* of the number while avoiding the need to write infinitely long decimals.

Writing rational numbers as fractions allows us to complete precise mathematical calculations easily with pen and paper, without the need for a calculator.

Example

Calculate the sum of $\frac{1}{7}$ and $\frac{1}{3}$.

Let's say we decide, for reasons unknown, that it's a great day for decimal notation—we'd have to write our answer as

$$\begin{aligned}
\text{ans} &= 0.\overline{142857} \;+\; 0.\overline{3}\\
&= 0.142\,857\,142\,857\ldots \;+\; 0.333\,333\,333\,333\ldots\\
&= 0.476\,190\,476\,190\,476\ldots\\
&= 0.4\overline{761904}.
\end{aligned}$$

Wow that was complicated! This calculation is much simpler if we use fractions:

$$\frac{1}{7}+\frac{1}{3}=\frac{3\times 1}{3\times 7}+\frac{1\times 7}{3\times 7}=\frac{3}{21}+\frac{7}{21}=\frac{3+7}{21}=\frac{10}{21}.$$

Definitions

The fraction "a over b" can be written in three different ways:

$$a/b \equiv a \div b \equiv \frac{a}{b}\,.$$

The top and bottom parts of a fraction have special names:

- b is called the *denominator* of the fraction. It tells us how many parts there are in the whole.
- a is called the *numerator* and it tells us the number of parts we are given.

Addition of fractions

Suppose we are asked to find the sum of the two fractions $\frac{a}{b}$ and $\frac{c}{d}$. If the denominators are the same, then we have to add just the top parts $\frac{1}{5}+\frac{2}{5}=\frac{3}{5}$. It makes sense to add the numerators since they refer to parts of the *same* whole.

However, if the denominators are different, we cannot add the numerators directly since they refer to "parts" from a different whole. Instead, we must rewrite the fractions so they have a *common denominator* before we can add the numerators. We can obtain a common denominator by multiplying the first fraction by $\frac{d}{d}=1$ and the second fraction by $\frac{c}{c}=1$ in order to make the denominator of both fractions the same:

$$\frac{a}{b}+\frac{c}{d}=\frac{a}{b}\left(\frac{d}{d}\right)+\left(\frac{b}{b}\right)\frac{c}{d}=\frac{ad}{bd}+\frac{bc}{bd}.$$

Now that we have fractions with the same denominator, we can add the numerators. The effect of multiplying the top and bottom of the fractions by the same number is the same as multiplying by 1 since the above operations did not change the fractions:

$$\frac{a}{b}+\frac{c}{d}=\frac{ad}{bd}+\frac{bc}{bd}=\frac{ad+bc}{bd}.$$

More generally, to add two fractions, we can pick the *least common multiple* $\mathrm{LCM}(b, d)$ to use as the common denominator. The LCM of two numbers is obtained by multiplying the numbers together and removing their common factors:

$$\mathrm{LCM}(b, d) = \frac{b \times d}{\mathrm{GCD}(b, d)}\,,$$

where $\mathrm{GCD}(b, d)$ is the *greatest common divisor* of b and d—the largest number that divides both b and d.

For example, to add $\frac{1}{6}$ and $\frac{1}{15}$, we rewrite the fractions so they have the common denominator 6×15 (the product of the two denominators). Or, more simply, we can see $6 = 3 \times 2$, and $15 = 3 \times 5$, meaning 3 is a common divisor of both 6 and 15. The least common multiple is then $\frac{6\times15}{3} = 30$, so we write:

$$\frac{1}{6} + \frac{1}{15} = \frac{5 \times 1}{5 \times 6} + \frac{1 \times 2}{15 \times 2} = \frac{5}{30} + \frac{2}{30} = \frac{7}{30}.$$

Actually, all this LCM and GCD business is not *required*—but it is the most efficient way to add fractions without having to deal with excessively large numbers. If you use the common denominator $b \times d$, you will arrive at the same answer as above after simplification:

$$\frac{1}{6} + \frac{1}{15} = \frac{15 \times 1}{15 \times 6} + \frac{1 \times 6}{15 \times 6} = \frac{15}{90} + \frac{6}{90} = \frac{21}{90} = \frac{7}{30}\,.$$

Multiplication of fractions

Fraction multiplication involves multiplying the numerators together and multiplying the denominators together:

$$\frac{a}{b} \times \frac{c}{d} = \frac{a \times c}{b \times d} = \frac{ac}{bd}\,.$$

Division of fractions

To divide two fractions, compute the product of the first fraction times the second fraction *flipped*:

$$\frac{a/b}{c/d} = \frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{a \times d}{b \times c} = \frac{ad}{bc}\,.$$

The *multiplicative inverse* of something times that something should give 1 as the answer. We obtain the multiplicative inverse of a fraction by interchanging the roles of the numerator and the denominator:

$$\left(\frac{c}{d}\right)^{-1} = \frac{d}{c}\,.$$

Any fraction times its multiplicative inverse gives $\frac{c}{d} \times \left(\frac{c}{d}\right)^{-1} = \frac{c}{d} \times \frac{d}{c} = \frac{cd}{cd} = 1$. The "flip and multiply" rule for division stems from the fact that division by a number x is the same as multiplication by $\frac{1}{x}$.

Whole and fraction notation

To indicate a fraction like $\frac{5}{3}$, which is greater than 1, we sometimes use the notation $1\frac{2}{3}$, which is read as "one and two thirds." Similarly, $\frac{22}{7} = 3\frac{1}{7}$.

There is nothing wrong with writing fractions like $\frac{5}{3}$ and $\frac{22}{7}$. However, some teachers say this way of writing fractions is *improper* and demand that fractions are written in the whole-and-fraction way, as in $1\frac{2}{3}$ and $3\frac{1}{7}$. At the end of the day, both notations are correct.

Repeating decimals

When written as decimal numbers, certain fractions have infinitely long decimal expansions. We use the overline notation to indicate the digit(s) that repeat in the expansion:

$$\frac{1}{3} = 0.\bar{3} = 0.333\ldots; \qquad \frac{1}{7} = 0.\overline{142857} = 0.14285714285714\ldots.$$

1.10 The number line

The number line is a useful graphical representation for numbers. The integers $\mathbb{Z}$ correspond to the notches on the line while the rationals $\mathbb{Q}$ and the reals $\mathbb{R}$ densely cover the whole line.

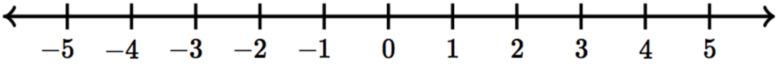


Figure 1.2: The number line is a visual representation of numbers.

You can see the *ordering* of the numbers from smallest on the left, to largest on the right. The line extends indefinitely on both sides: the left side goes all the way to negative infinity $-\infty$ and the right side stretches to positive infinity ∞.

Definitions

We use the following notation to denote subsets of the real line:

- $x \in I$: indicates the variable x lies in the interval I. The expression reads "x is an element of I," or simply "x is in I."
- $[a, b]$: the *closed* interval from a to b. This corresponds to the set of numbers between a and b on the real line, including the endpoints a and b. $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$.
- (a, b): the *open* interval from a to b. This corresponds to the set of numbers between a and b on the real line, *not* including the endpoints a and b. $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$.
- $[a, b)$: the mixed interval that includes the left endpoint a but not the right endpoint b. $[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$.

Sometimes we encounter intervals that consist of two disjointed parts. We use the notation $[a, b] \cup [c, d]$ to denote the set of all numbers found *either* between a and b (inclusive) *or* between c and d (inclusive).

Intervals

We can graphically represent *intervals* of the real numbers by setting a section of the number line in **bold**. For example, the set of numbers that are strictly greater than 2 and strictly smaller than 4 is represented mathematically either as "$(2, 4)$" or more explicitly as

$$\{x \in \mathbb{R} \mid 2 < x < 4\}.$$

If this is the first time you've seen the formal definition of a *set*, don't be alarmed. Let's parse this mathematical expression together. The symbol $\in$ denotes set membership. The vertical bar stands for "such that." The whole expression "$\{x \in \mathbb{R} \mid 2 < x < 4\}$" is read "the set of real numbers x, such that $2 < x < 4$."

The graphical representation of this interval is shown in Figure 1.3.

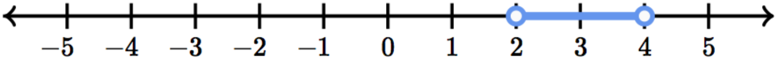


Figure 1.3: The open interval $(2, 4) \equiv \{x \in \mathbb{R} \mid 2 < x < 4\}$.

Note that this subset is described by *strict* inequalities, which means the subset contains 2.000000001 and 3.99999999, but doesn't contain the endpoints 2 and 4. These *open* endpoints 2 and 4 are denoted on the number line as empty dots. An empty dot indicates that the endpoint is not included in the set.

We use the *union* symbol ($\cup$) to denote subsets of the number line that consist of several parts. For example, the set of numbers that lies *either* between -3 and 0 *or* between 1 and 2 is written as

$$\{x \in \mathbb{R} \mid -3 \leq x \leq 0\} \ \cup \ \{x \in \mathbb{R} \mid 1 \leq x \leq 2\}.$$

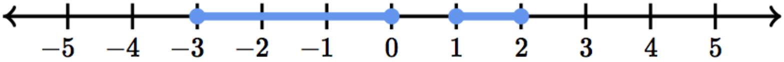


Figure 1.4: The graphical representation of the set $[-3, 0] \cup [1, 2]$.

This set is defined by less-than-or-equal inequalities, so the intervals contain their endpoints. These *closed* endpoints are denoted on the number line with filled-in dots.

1.11 Inequalities

To solve an equation, we need to find the one (or many) values of x that satisfy the equation. An equation's *solution set* consists of a discrete set of values. For example, the solutions to $(x-3)^2 = 4$ are $x = 1$ and $x = 5$.

In this section, we'll learn how to solve equations that involve inequalities. The solution to an inequality is usually denoted as an *interval*—a subset of the number line. For example, the inequality $(x-3)^2 \leq 4$ is equivalent to asking the question, "For which values of x is $(x-3)^2$ less than or equal to 4?" The answer is the interval $[1, 5] \equiv \{x \in \mathbb{R} \mid 1 \leq x \leq 5\}$.

We approach inequalities with roughly the same techniques we learned for dealing with equations: we must perform simplifying steps **on both sides of the inequality** until we obtain the answer.

Definitions

The different types of inequality conditions are:

- $f(x) < g(x)$: a strict inequality. The function $f(x)$ is always *strictly less than* $g(x)$.
- $f(x) \leq g(x)$: the function $f(x)$ is *less than or equal to* the function $g(x)$.
- $f(x) > g(x)$: $f(x)$ is *strictly greater than* $g(x)$.
- $f(x) \geq g(x)$: $f(x)$ is *greater than or equal to* $g(x)$.

The solutions to an inequality correspond to subsets of the real line. Depending on the type of inequality, the answer will be either a *closed* or *open* interval.

Formulas

The main idea for solving inequalities is the same as the main idea for solving equations, except for one small, special step. When multiplying by a negative number on both sides, the direction of the inequality must be flipped:

$$f(x) \leq g(x) \qquad \Rightarrow \qquad -f(x) \geq -g(x).$$

Example To solve $(x-3)^2 \leq 4$ we must *dig* toward the x and *undo* all the operations that stand in our way:

$$\begin{aligned} (x-3)^2 &\leq 4, \\ -2 \leq (x-3) &\leq 2, \\ 1 \leq \quad x \quad &\leq 5. \end{aligned}$$

In the first step, we took the square root operation (the inverse of the quadratic function), then added 3 to both sides. The final answer is $x \in [1,5]$ or, written more explicitly, $\{x \in \mathbb{R} \mid 1 \leq x \leq 5\}$.

Discussion

Solving inequalities is no more complicated than solving equations. You can think about an inequality in terms of its end points, which correspond to the equality condition.

1.12 The Cartesian plane

Named after famous philosopher and mathematician René Descartes, the Cartesian plane is a graphical representation for *pairs* of numbers.

Generally, we call the plane's horizontal axis "the x-axis" and its vertical axis "the y-axis." We put notches at regular intervals on each axis so we can measure distances. Figure 1.5 is an example of an empty Cartesian coordinate system. Think of the coordinate system as an empty canvas. What can you draw on this canvas?

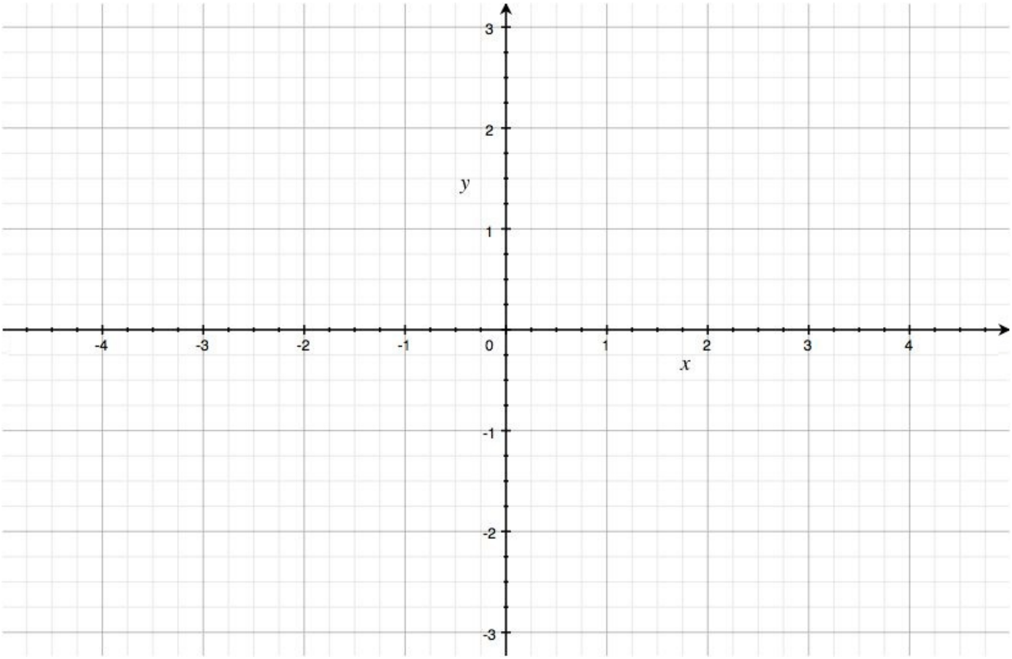


Figure 1.5: The (x, y)-coordinate system, which is also known as the Cartesian plane. Points $P = (P_x, P_y)$, vectors $\vec{v} = (v_x, v_y)$, and graphs of functions $(x, f(x))$ live here.

Vectors and points

A *point* $P = (P_x, P_y)$ in the Cartesian plane has an x-coordinate and a y-coordinate. To find this point, start from the origin—the point (0,0)—and move a distance P_x on the x-axis, then move a distance P_y on the y-axis.

Similar to a point, a vector $\vec{v} = (v_x, v_y)$ is a pair of coordinates. Unlike points, we don't necessarily start from the plane's origin when mapping vectors. We draw vectors as arrows that explicitly mark where the vector starts and where it ends. Note that vectors $\vec{v}_2$ and $\vec{v}_3$ illustrated in Figure 1.6 are actually the *same* vector—the "displace left by 1 and down by 2" vector. It doesn't matter where you draw this vector, it will always be the same whether it begins at the plane's origin or elsewhere.

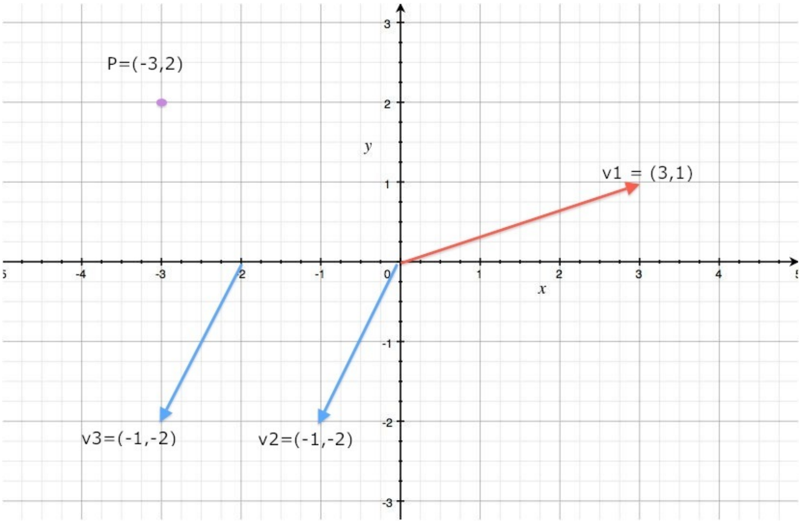


Figure 1.6: A Cartesian plane which shows the point $P = (-3, 2)$ and the vectors $\vec{v}_1 = (3, 1)$ and $\vec{v}_2 = \vec{v}_3 = (-2, -1)$.

Graphs of functions

The Cartesian plane is great for visualizing functions,

$$f : \mathbb{R} \to \mathbb{R}.$$

You can think of a function as a set of input-output pairs $(x, f(x))$. You can *graph* a function by letting the y-coordinate represent the function's output value:

$$(x, y) = (x, f(x)).$$

For example, with the function $f(x) = x^2$, we can pass a line through the set of points

$$(x, y) = (x, x^2),$$

and obtain the graph shown in Figure 1.7.

When plotting functions by setting $y = f(x)$, we use a special terminology for the two axes. The x-axis represents the *independent* variable (the one that varies freely), and the y-axis represents the *dependent* variable $f(x)$, since $f(x)$ depends on x.

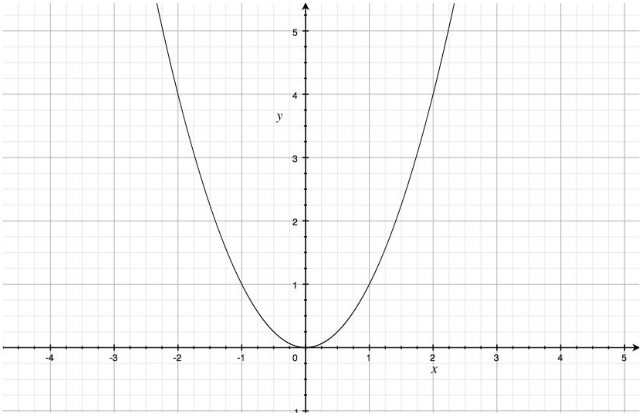


Figure 1.7: The graph of the function $f(x) = x^2$ consists of all pairs of points (x, y) in the Cartesian plane that satisfy $y = x^2$.

To draw the graph of any function $f(x)$, use the following procedure. Imagine making a sweep over all of the possible input values for the function. For each input x, put a point at the coordinates $(x, y) = (x, f(x))$ in the Cartesian plane. Using the graph of a function, you can literally *see* what the function does: the "height" y of the graph at a given x-coordinate tells you the value of the function $f(x)$.

Discussion

To build mathematical intuition, it is essential you understand the graphs of functions. Trying to memorize the definitions and the properties of functions is a difficult task. Remembering what the function "looks like" is comparatively easier. You should spend some time familiarizing yourself with the graphs of the functions presented in the next section.

1.13 Functions

We need to have a relationship talk. We need to talk about functions. We use functions to describe the relationships between variables. In particular, functions describe how one variable *depends* on another.

For example, the revenue R from a music concert depends on the number of tickets sold n. If each ticket costs $\$25$, the revenue from the concert can be written *as a function of* n as follows: $R(n) = 25n$. Solving for n in the equation $R(n) = 7000$ tells us the number of ticket sales needed to generate $\$7000$ in revenue. This is a simple model of a function; as your knowledge of functions builds, you'll learn how to build more detailed models of reality. For instance, if you need to include a 5% processing charge for issuing the tickets, you can update the revenue model to $R(n) = 0.95 \cdot 25 \cdot n$. If the estimated cost of hosting the concert is $C = \$2000$, then the profit from the concert P can be modelled as

$$\begin{aligned} P(n) &= R(n) \ - \ C \\ &= 0.95 \cdot \$25 \cdot n \ - \ \$2000 \end{aligned}$$

The function $P(n) = 23.75n - 2000$ models the profit from the concert as a function of the number of tickets sold. This is a pretty good model already, and you can always update it later on as you find out more information.

The more functions you know, the more tools you have for modelling reality. To "know" a function, you must be able to understand and connect several of its aspects. First you need to know the function's mathematical **definition**, which describes exactly what the function does. Starting from the function's definition, you can use your existing math skills to find the function's domain, its range, and its inverse function. You must also know the **graph** of the function; what the function looks like if you plot x versus $f(x)$ in the Cartesian plane. It's also a good idea to remember the **values** of the function for some important inputs. Finally—and this is the part that takes time—you must learn about the function's **relations** to other functions.

Definitions

A *function* is a mathematical object that takes numbers as inputs and gives numbers as outputs. We use the notation

$$f\colon A \to B$$

to denote a function from the input set A to the output set B. In this book, we mostly study functions that take real numbers as inputs and give real numbers as outputs: $f\colon \mathbb{R} \to \mathbb{R}$.

We now define some fancy technical terms used to describe the input and output sets.

- The *domain* of a function is the set of allowed input values.
- The *image* or *range* of the function f is the set of all possible output values of the function.
- The *codomain* of a function describes the type of outputs the function has.

To illustrate the subtle difference between the image of a function and its codomain, consider the function $f(x) = x^2$. The quadratic function is of the form $f\colon \mathbb{R} \to \mathbb{R}$. The function's domain is $\mathbb{R}$ (it takes real numbers as inputs) and its codomain is $\mathbb{R}$ (the outputs are real numbers too), however, not all outputs are possible. The *image* of the function $f(x) = x^2$ consists only of the nonnegative real numbers $[0, \infty \equiv \{y \in \mathbb{R} \mid y \geq 0\}$.

A function is not a number; rather, it is a *mapping* from numbers to numbers. For any input x, the output value of f for that input is denoted $f(x)$.

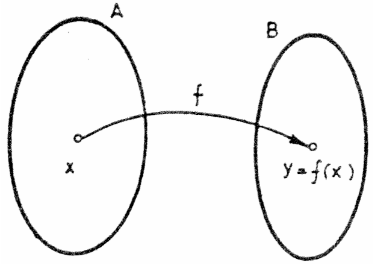


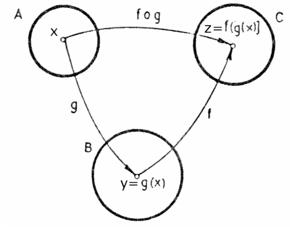
Figure 1.8: An abstract representation of a function f from the set A to the set B. The function f is the arrow which *maps* each input x in A to an output $f(x)$ in B. The output of the function $f(x)$ is also denoted y.

We say "f maps x to $f(x)$," and use the following terminology to classify the type of mapping that a function performs:

- A function is *one-to-one* or *injective* if it maps different inputs to different outputs.
- A function is *onto* or *surjective* if it covers the entire output set (in other words, if the image of the function is equal to the function's codomain).
- A function is *bijective* if it is both injective and surjective. In this case, f is a *one-to-one correspondence* between the input set and the output set: for each of the possible outputs $y \in Y$ (surjective part), there exists exactly one input $x \in X$, such that $f(x) = y$ (injective part).

The term *injective* is an allusion from the 1940s inviting us to picture the actions of injective functions as pipes through which numbers flow like fluids. Since a fluid cannot be compressed, the output space must be at least as large as the input space. A modern synonym for injective functions is to say they are *two-to-two*. If we imagine two specks of paint floating around in the "input fluid," an injective function will contain two distinct specks of paint in the "output fluid." In contrast, non-injective functions can map several different inputs to the same output. For example $f(x) = x^2$ is not injective since the inputs 2 and -2 are both mapped to the output value 4.

Function composition

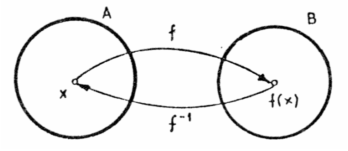


We can combine two simple functions by chaining them together to build a more complicated function. This act of applying one function after another is called *function composition*. Consider for example the composition:

$$f \circ g\,(x) \equiv f(\;g(x)\;) = z.$$

The diagram on the right illustrates what is going on. First, the function $g : A \to B$ acts on some input x to produce an intermediary value $y = g(x)$ in the set B. The intermediary value y is then passed through the function $f : B \to C$ to produce the final output value $z = f(y) = f(g(x))$ in the set C. We can think of the *composite function* $f \circ g$ as a function in its own right. The function $f \circ g : A \to C$ is defined through the formula $f \circ g\,(x) \equiv f(g(x))$.

Inverse function



Recall that a *bijective* function is a one-to-one correspondence between a set of input values and a set of output values. Given a bijective function $f : A \to B$, there exists an inverse function $f^{-1} : B \to A$, which performs the *inverse mapping* of f. If you start from some x, apply f, and then apply f^{-1}, you'll arrive—full circle—back to the original input x:

$$f^{-1}\big(\;f(x)\;\big) \equiv f^{-1} \circ f\;(x) = x.$$

This inverse function is represented abstractly as a backward arrow, that puts the value $f(x)$ back to the x it came from.

Function names

We use short symbols like $+$, $-$, $\times$, and $\div$ to denote most of the important functions used in everyday life. We also use the weird *surd* notation to denote n^{th}

root $\sqrt[n]{\ }$ and the superscript notation to denote exponents. All other functions are identified and denoted by their *name*. If I want to compute the *cosine* of the angle $60°$ (a function describing the ratio between the length of one side of a right-angle triangle and the hypotenuse), I write $\cos(60°)$, which means I want the value of the $\cos$ function for the input $60°$.

Incidentally, the function $\cos$ has a nice output value for that specific angle: $\cos(60°) \equiv \frac{1}{2}$. Therefore, seeing $\cos(60°)$ somewhere in an equation is the same as seeing $\frac{1}{2}$. To find other values of the function, say $\cos(33.13°)$, you'll need a calculator. A scientific calculator features a convenient little $\boxed{\cos}$ button for this very purpose.

Handles on functions

When you learn about functions you learn about the different "handles" by which you can "grab" these mathematical objects. The main handle for a function is its **definition**: it tells you the precise way to calculate the output when you know the input. The function definition is an important handle, but it is also important to "feel" what the function does intuitively. How does one get a feel for a function?

Table of values

One simple way to represent a function is to look at a list of input-output pairs: $\big\{\{\text{in} = x_1, \text{out} = f(x_1)\}, \{\text{in} = x_2, \text{out} = f(x_2)\}, \{\text{in} = x_3, \text{out} = f(x_3)\}, \ldots\big\}$. A more compact notation for the input-output pairs is $\{(x_1, f(x_1)), (x_2, f(x_2)), (x_3, f(x_3)), \ldots\}$. You can make your own little **table of values**, pick some random inputs, and record the output of the function in the second column:

$$\begin{array}{rcl}
\text{input} = x & \rightarrow & f(x) = \text{output} \\
0 & \rightarrow & f(0) \\
1 & \rightarrow & f(1) \\
55 & \rightarrow & f(55) \\
x_4 & \rightarrow & f(x_4).
\end{array}$$

In addition to choosing random numbers for your table, it's also generally a good idea to check the function's values at $x = 0$, $x = 1$, $x = 100$, $x = -1$, and any other important-looking x value.

Function graph

One of the best ways to feel a function is to look at its graph. A graph is a line on a piece of paper that passes through all input-output pairs of a function. Imagine you have a piece of paper, and on it you draw a blank *coordinate system* as in Figure 1.9.

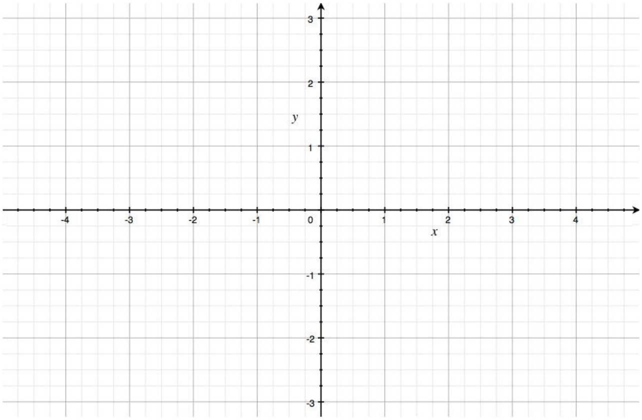


Figure 1.9: An empty (x, y)-coordinate system that you can use to plot the graph of *any* function $f(x)$. The graph of $f(x)$ consists of all the points for which $(x, y) = (x, f(x))$. See Figure 1.7 on page 49 for the graph of $f(x) = x^2$.

The horizontal axis, sometimes called the *abscissa*, is used to measure x. The vertical axis is used to measure $f(x)$. Because writing out $f(x)$ every time is long and tedious, we use a short, single-letter alias to denote the output value of f as follows:

$$y \equiv f(x) = \text{output}.$$

Think of each input-output pair of the function f as a point (x, y) in the coordinate system. The graph of a function is a representational drawing of everything the function does. If you understand how to interpret this drawing, you can infer everything there is to know about the function.

Facts and properties

Another way to feel a function is by knowing the function's properties. This approach boils down to learning facts about the function and its relation to other functions. An example of a mathematical fact is $\sin(30°) = \frac{1}{2}$. An example of a mathematical relation is the equation $\sin^2 x + \cos^2 x = 1$, which indicates a link between the $\sin$ function and the $\cos$ function.

The more you know about a function, the more "paths" your brain builds to connect to that function. Real math knowledge is not memorization; it requires establishing a graph of associations between different areas of information in your brain. Each concept is a *node* in this graph, and each fact you know about this concept is an *edge*. Mathematical thought is the usage of this graph to produce calculations and mathematical arguments called proofs. For example, by connecting your knowledge of the fact $\sin(30°) = \frac{1}{2}$ with the relation $\sin^2 x + \cos^2 x = 1$, you can show that $\cos(30°) = \frac{\sqrt{3}}{2}$. Note the notation $\sin^2(x)$ means $(\sin(x))^2$.

To develop mathematical skills, it is vital to practice path-building between related concepts by solving exercises and reading and writing mathematical proofs. With this book, I will introduce you to many paths between concepts; it's up to you to reinforce these by using what you've learned to solve problems.

Example

Consider the function f from the real numbers to the real numbers ($f\colon \mathbb{R} \to \mathbb{R}$) defined by the quadratic expression

$$f(x) = x^2 + 2x + 3.$$

The value of f when $x = 1$ is $f(1) = 1^2 + 2(1) + 3 = 1 + 2 + 3 = 6$. When $x = 2$, the output is $f(2) = 2^2 + 2(2) + 3 = 4 + 4 + 3 = 11$. What is the value of f when $x = 0$?

Example 2

Consider the exponential function with base 2:

$$f(x) = 2^x.$$

This function is crucial to computer systems. For instance, RAM memory chips come in powers of two because the memory space is exponential in the number of "address lines" used on the chip. When $x = 1$, $f(1) = 2^1 = 2$. When x is 2 we have $f(2) = 2^2 = 4$. The function is therefore described by the following input-output pairs: $(0, 1)$, $(1, 2)$, $(2, 4)$, $(3, 8)$, $(4, 16)$, $(5, 32)$, $(6, 64)$, $(7, 128)$, $(8, 256)$, $(9, 512)$, $(10, 1024)$, $(11, 2048)$, $(12, 4096)$, etc. Recall that any number raised to exponent 0 gives 1. Thus, the exponential function passes through the point $(0, 1)$. Recall also that negative exponents lead to fractions: $(-1, \frac{1}{2^1} = \frac{1}{2}), (-2, \frac{1}{2^2} = \frac{1}{4}), (-3, \frac{1}{2^3} = \frac{1}{8})$, etc.

Discussion

In this section we talked a lot about functions in general, but we haven't said much about any function specifically. There are many useful functions out there, and we can't discuss them all here. In the next section, we'll introduce 10 functions of strategic importance for all of science. If you get a grip on these functions, you'll be able to understand all of physics and calculus and handle *any* problem your teacher may throw at you.

Links

[Tank game where you specify the function of the projectile]
http://www.graphwar.com/play.html

1.14 Function reference

Your *function vocabulary* determines how well you can express yourself mathematically in the same way that your English vocabulary determines how well you can express yourself in English. The following pages aim to embiggen your function vocabulary so you won't be caught with your pants down when the teacher tries to pull some trick on you at the final.

If you are seeing these functions for the first time, don't worry about remembering all the facts and properties on the first reading. We will use these functions throughout the rest of the book so you will have plenty of time to become familiar with them. Just remember to come back to this section if you ever get stuck on a function.

Line

The equation of a line describes an input-output relationship where the change in the output is *proportional* to the change in the input. The equation of a line is

$$f(x) = mx + b.$$

The constant m describes the slope of the line. The constant b is called the y-intercept and it corresponds to the value of the function when $x = 0$.

Graph

Properties

- Domain: $x \in \mathbb{R}$.
  The function $f(x) = mx + b$ is defined for all input values $x \in \mathbb{R}$.
- Image: $x \in \mathbb{R}$ if $m \neq 0$. If $m = 0$ the function is constant $f(x) = b$, so the image set contains only a single number $\{b\}$.
- b/m: the x-intercept of $f(x) = mx + b$. The x-intercept is obtained by solving $f(x) = 0$.

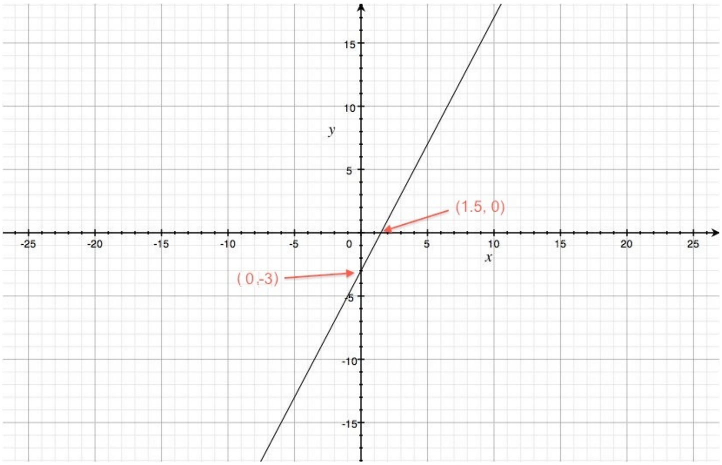


Figure 1.10: The graph of the function $f(x) = 2x - 3$. The slope is $m = 2$. The y-intercept of this line is at $y = -3$. The x-intercept is at $x = \frac{3}{2}$.

- A unique line passes through any two points (x_1, y_1) and (x_2, y_2) if $x_1 \neq x_2$.
- The inverse to the line $f(x) = mx + b$ is $f^{-1}(x) = \frac{1}{m}(x - b)$, which is also a line.

General equation

A line can also be described in a more symmetric form as a relation:

$$Ax + By = C.$$

This is known as the *general* equation of a line. The general equation for the line shown in Figure 1.10 is $2x - 1y = 3$.

Given the general equation of a line $Ax + By = C$, you can convert to the function form $y = f(x) = mx + b$ using $b = \frac{C}{B}$ and $m = \frac{-A}{B}$.

Square

The function x *squared*, is also called the *quadratic* function, or *parabola*. The formula for the quadratic function is

$$f(x) = x^2.$$

The name "quadratic" comes from the Latin *quadratus* for square, since the expression for the area of a square with side length x is x^2.

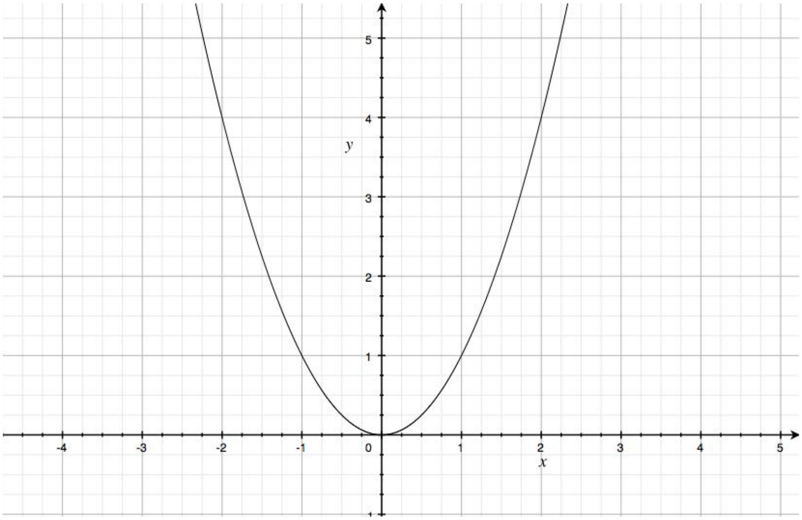


Figure 1.11: Plot of the quadratic function $f(x) = x^2$. The graph of the function passes through the following (x, y) coordinates: $(-2, 4)$, $(-1, 1)$, $(0, 0)$, $(1, 1)$, $(2, 4)$, $(3, 9)$, etc.

Properties

- Domain: $x \in \mathbb{R}$.
  The function $f(x) = x^2$ is defined for all input values $x \in \mathbb{R}$.
- Image: $f(x) \in [0, \infty$.
  The outputs are never negative: $x^2 \geq 0$, for all $x \in \mathbb{R}$.

- The function x^2 is the inverse of the square root function $\sqrt{x}$.
- $f(x) = x^2$ is *two-to-one*: it sends both x and $-x$ to the same output value $x^2 = (-x)^2$.
- The quadratic function is *convex*, meaning it curves upward.

Square root

The square root function is defined as

$$f(x) = \sqrt{x} \equiv x^{\frac{1}{2}}.$$

The square root $\sqrt{x}$ is the inverse function of the quadratic x^2.

Graph

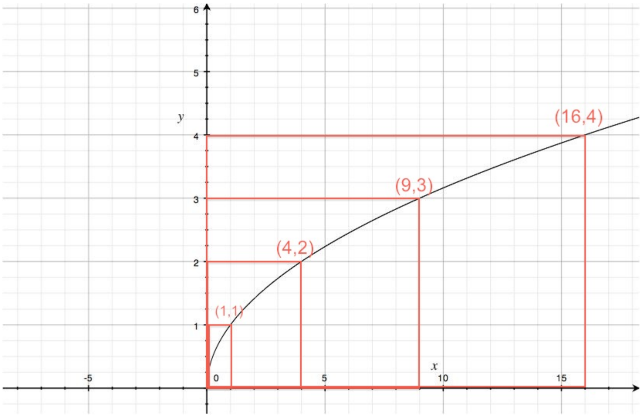


Figure 1.12: The graph of the function $f(x) = \sqrt{x}$. The domain of the function is $x \in [0, \infty$. You can't take the square root of a negative number.

Properties

- Domain: $x \in [0, \infty$.
  The function $f(x) = \sqrt{x}$ is only defined for nonnegative inputs $x \geq 0$. There

is no real number y such that y^2 is negative, hence the function $f(x) = \sqrt{x}$ is not defined for negative inputs x.

- Image: $f(x) \in [0, \infty$.
  The outputs of the function $f(x) = \sqrt{x}$ are never negative: $\sqrt{x} \geq 0$, for all $x \in [0, \infty$.

In addition to *square* root, there is also *cube* root $f(x) = \sqrt[3]{x} \equiv x^{\frac{1}{3}}$, which is the inverse function for the cubic function $f(x) = x^3$. We have $\sqrt[3]{8} = 2$ since $2 \times 2 \times 2 = 8$. More generally, we can define the root n^{th} root function $\sqrt[n]{x}$ as the inverse function of x^n.

Absolute value

The absolute value function tells us the *size* of numbers without paying attention to whether the number is positive or negative. We can compute a number's absolute value by *ignoring the sign* of the input number. Thus, a number's absolute value corresponds to its distance from the origin of the number line.

Another way of thinking about the absolute value function is to say it multiplies negative numbers by -1 to "cancel" their negative sign:

$$f(x) = |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Properties

- Always returns a non-negative number
- The combination of squaring followed by square-root is equivalent to the absolute value function:
  $$\sqrt{x^2} \equiv |x|,$$
  since squaring destroys the sign.

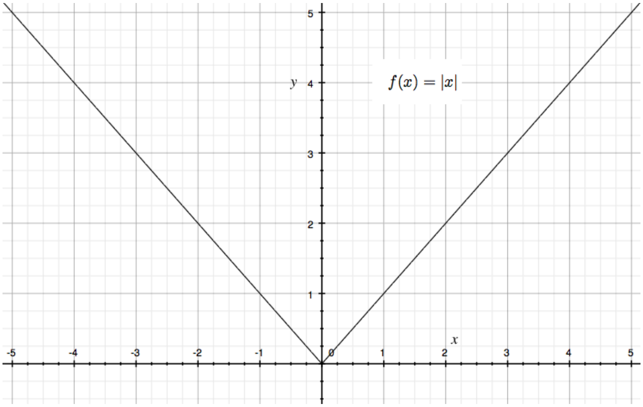


Figure 1.13: The graph of the absolute value function $f(x) = |x|$.

Polynomial functions

The general equation for a polynomial function of degree n is written,

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n.$$

The constants a_i are known as the *coefficients* of the polynomial.

Parameters

- n: the *degree* of the polynomial
- a_0: the constant term
- a_1: the *linear* coefficient, or first-order coefficient
- a_2: the *quadratic* coefficient
- a_3: the *cubic* coefficient
- a_n: the n^{th} order coefficient

A polynomial of degree n has $n+1$ coefficients: $a_0, a_1, a_2, \ldots, a_n$.

Properties

- Domain: $x \in \mathbb{R}$. Polynomials are defined for all inputs $x \in \mathbb{R}$.
- Image: depends on the coefficients
- The sum of two polynomials is also a polynomial.

Even and odd functions

The polynomials form an entire family of functions. Depending on the choice of degree n and coefficients a_0, a_1, $\ldots$, a_n, a polynomial function can take on many different shapes. We will study polynomials and their properties in more detail in Section 1.15, but for now consider the following observations about the symmetries of polynomials:

- If a polynomial contains only even powers of x, like $f(x) = 1 + x^2 - x^4$ for example, we call this polynomial *even*. Even polynomials have the property $f(x) = f(-x)$. The sign of the input doesn't matter.
- If a polynomial contains only odd powers of x, for example $g(x) = x+x^3-x^9$, we call this polynomial *odd*. Odd polynomials have the property $g(x) = -g(-x)$.
- If a polynomial has both even and odd terms then it is neither even nor odd.

Note that the terminology of *odd* and *even* applies to functions in general and not just to polynomials. All functions which satisfy $f(x) = f(-x)$ are called even, and all functions which satisfy $f(x) = -f(-x)$ are called odd.

Sine

The sine function represents a fundamental unit of vibration. The graph of $\sin(x)$ *oscillates* up and down and crosses the x-axis multiple times. The shape of the graph of $\sin(x)$ corresponds to the shape of a vibrating string. See Figure 1.14.

In the remainder of this book, we'll meet the function $\sin(x)$ many times. We will define the function $\sin(x)$ more formally as a trigonometric ratio in Section 1.16. In Chapter 3 we will use $\sin(x)$ and $\cos(x)$ (another trigonometric ratio) to work out the *components* of vectors. Later in Chapter 4, we will learn how the sine function can be used to describe waves and periodic motion.

At this point in the book, however, we don't want to go into too much detail about all these applications. Let's hold off the discussion about vectors, triangles, angles, and ratios of lengths of sides and instead just focus on the graph of the function $f(x) = \sin(x)$.

Graph

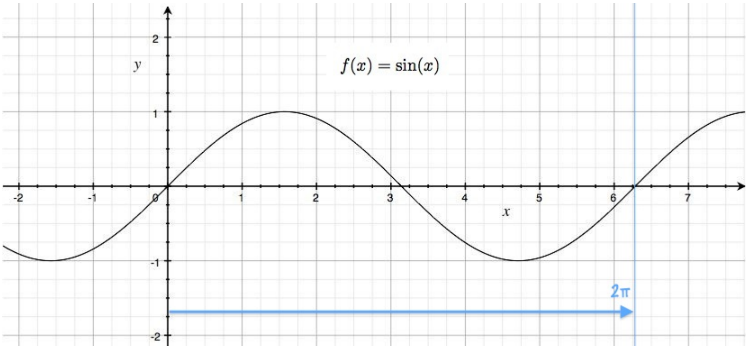
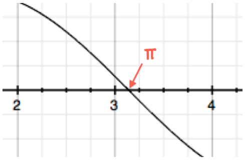


Figure 1.14: The graph of the function $y = \sin(x)$ passes through the following (x, y) coordinates: $(0, 0)$, $(\frac{\pi}{6}, \frac{1}{2})$, $(\frac{\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{\pi}{3}, \frac{\sqrt{3}}{2})$, $(\frac{\pi}{2}, 1)$, $(\frac{2\pi}{3}, \frac{\sqrt{3}}{2})$, $(\frac{3\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{5\pi}{6}, \frac{1}{2})$, and $(\pi, 0)$. For $x \in [\pi, 2\pi]$ the function has the same shape as for $x \in [0, \pi]$ but with negative values.



Let's start at $x = 0$ and follow the graph of the function $\sin(x)$ as it goes up and down. The graph starts from $(0, 0)$ and smoothly increases until it reaches the maximum value at $x = \frac{\pi}{2}$. Afterward, the function comes back down to cross the x-axis at $x = \pi$. After π, the function drops below the x-axis and reaches its minimum value of -1 at $x = \frac{3\pi}{2}$. It then travels up again to cross the x-axis at $x = 2\pi$. This 2π-long cycle repeats after $x = 2\pi$. This is why we call the function *periodic*—the shape of the graph repeats.

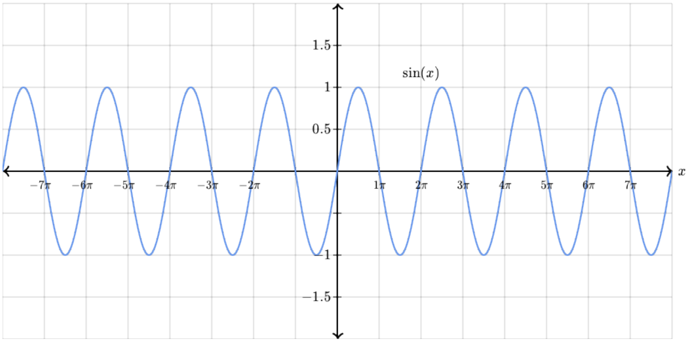


Figure 1.15: The graph of $\sin(x)$ from $x = 0$ to $x = 2\pi$ repeats periodically everywhere else on the number line.

Properties

- Domain: $x \in \mathbb{R}$.
  The function $f(x) = \sin(x)$ is defined for all input values $x \in \mathbb{R}$.
- Image: $\sin(x) \in [-1, 1]$.
  The outputs of the sine function are always between -1 and 1.
- Roots: $[\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi, \ldots]$.
  The function $\sin(x)$ has roots at all multiples of π.
- The function is periodic, with period 2π: $\sin(x) = \sin(x + 2\pi)$.
- The $\sin$ function is *odd*: $\sin(x) = -\sin(-x)$.
- Relation to $\cos$: $\sin^2 x + \cos^2 x = 1$
- Relation to $\csc$: $\csc(x) \equiv \frac{1}{\sin x}$ ($\csc$ is read *cosecant*)
- The inverse function of $\sin(x)$ is denoted as $\sin^{-1}(x)$, not to be confused with $(\sin(x))^{-1} = \frac{1}{\sin(x)} \equiv \csc(x)$. Sometimes the function $\sin^{-1}(x)$ is denoted "$\arcsin(x)$."

Cosine

The cosine function is the same as the sine function *shifted* a length $\frac{\pi}{2}$ to the left: $\cos(x) = \sin(x + \frac{\pi}{2})$. Thus everything you learned about the sine function also applies to the cosine function.

Graph

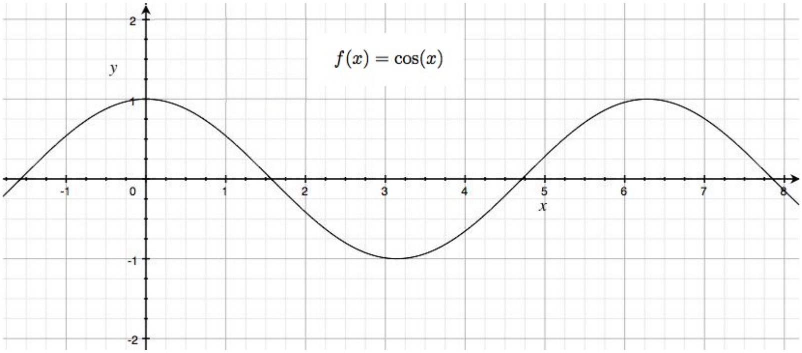


Figure 1.16: The graph of the function $y = \cos(x)$ passes through the following (x, y) coordinates: $(0, 1)$, $(\frac{\pi}{6}, \frac{\sqrt{3}}{2})$, $(\frac{\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{\pi}{3}, \frac{1}{2})$, $(\frac{\pi}{2}, 0)$, $(\frac{2\pi}{3}, -\frac{1}{2})$, $(\frac{3\pi}{4}, -\frac{\sqrt{2}}{2})$, $(\frac{5\pi}{6}, -\frac{\sqrt{3}}{2})$, and $(\pi, -1)$.

The cos function starts at $\cos(0) = 1$, then drops down to cross the x-axis at $x = \frac{\pi}{2}$. Cos continues until it reaches its minimum value at $x = \pi$. The function then moves upward, crossing the x-axis again at $x = \frac{3\pi}{2}$, and reaching its maximum value at $x = 2\pi$.

Properties

- Domain: $x \in \mathbb{R}$
- Image: $\cos(x) \in [-1, 1]$
- Relation to $\sin$: $\sin^2 x + \cos^2 x = 1$

- Relation to sec: $\sec(x) \equiv \frac{1}{\cos x}$ (sec is read *secant*)
- The inverse function of $\cos(x)$ is denoted $\cos^{-1}(x)$.
- The cos function is *even*: $\cos(x) = \cos(-x)$.

Tangent

The tangent function is the ratio of the sine and cosine functions:

$$f(x) = \tan(x) \equiv \frac{\sin(x)}{\cos(x)}.$$

Graph

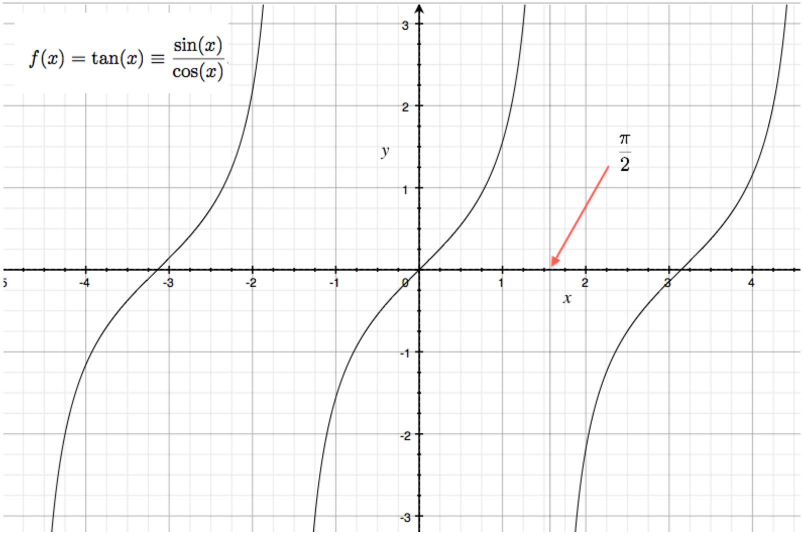


Figure 1.17: The graph of the function $f(x) = \tan(x)$.

Properties

- Domain: $\{x \in \mathbb{R} \mid x \neq \frac{(2n+1)\pi}{2} \text{ for any } n \in \mathbb{Z}\}$.

- Range: $x \in \mathbb{R}$.
- The function $\tan$ is periodic with period π (unlike $\sin$ and $\cos$, which are periodic with period 2π).
- The $\tan$ function "blows up" at all values of x for which the denominator $(\cos)$ goes to zero. These locations are called the *asymptotes* of the function. Their locations are $x = \dots, \frac{-3\pi}{2}, \frac{-\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \dots$.
- Value at 0: $\tan(0) = \frac{0}{1} = 0$, because $\sin(0) = 0$.
- The angle $x = \frac{\pi}{4}$ is special since both $\sin$ and $\cos$ are equal:

$$\tan\left(\frac{\pi}{4}\right) = \frac{\sin\left(\frac{\pi}{4}\right)}{\cos\left(\frac{\pi}{4}\right)} = \frac{\frac{\sqrt{2}}{2}}{\frac{\sqrt{2}}{2}} = 1.$$

Exponential

The exponential function base $e = 2.7182818\dots$ is denoted

$$f(x) = e^x \equiv \exp(x).$$

Properties

- Domain: $x \in \mathbb{R}$
- Range: $e^x \in (0, \infty$
- $f(a)f(b) = f(a+b)$ since $e^a e^b = e^{a+b}$.
- The derivative (the slope of the graph) of the exponential function is equal to the exponential function: $f(x) = e^x \quad \Rightarrow \quad f'(x) = e^x$.

A more general exponential function would be $f(x) = Ae^{\gamma x}$, where A is the initial value, and γ (the Greek letter *gamma*) is the *rate* of the exponential. For $\gamma > 0$, the function $f(x)$ is increasing, as in Figure 1.18. For $\gamma < 0$, the function is decreasing and tends to zero for large values of x. The case $\gamma = 0$ is special since $e^0 = 1$, so $f(x)$ is a constant of $f(x) = A1^x = A$.

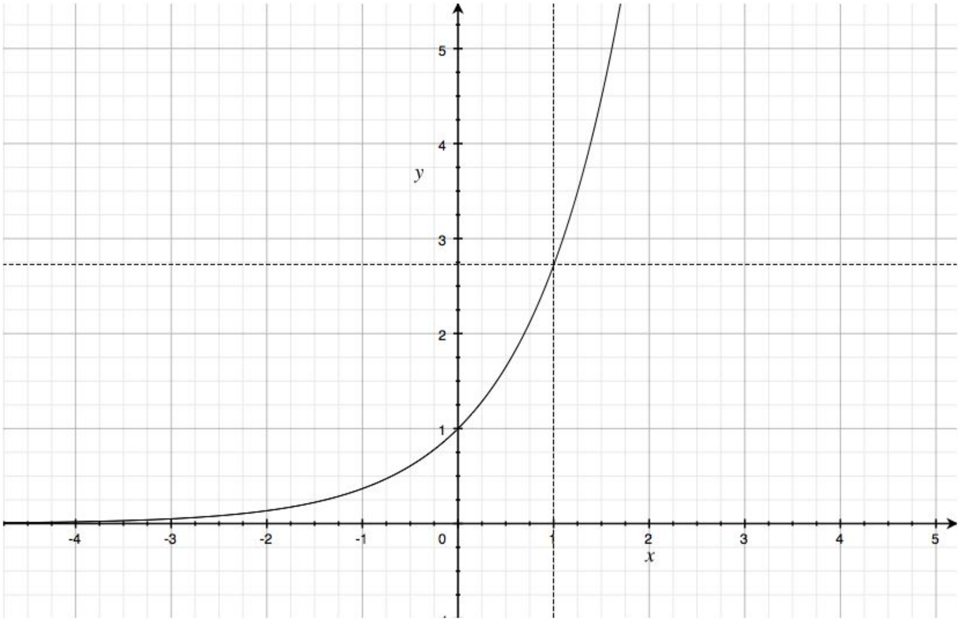


Figure 1.18: The graph of the exponential function $f(x) = e^x$ passes through the following (x, y) coordinates: $(-2, \frac{1}{e^2})$, $(-1, \frac{1}{e})$, $(0, 1)$, $(1, e)$, $(2, e^2)$, $(3, e^3 = 20.08\ldots)$, $(5, 148.41\ldots)$, and $(10, 22026.46\ldots)$.

Links

[The exponential function 2^x evaluated]
http://www.youtube.com/watch?v=e4MSN6IImpI

Natural logarithm

The natural logarithm function is denoted

$$f(x) = \ln(x) = \log_e(x).$$

The function $\ln(x)$ is the inverse function of the exponential e^x.

Graph

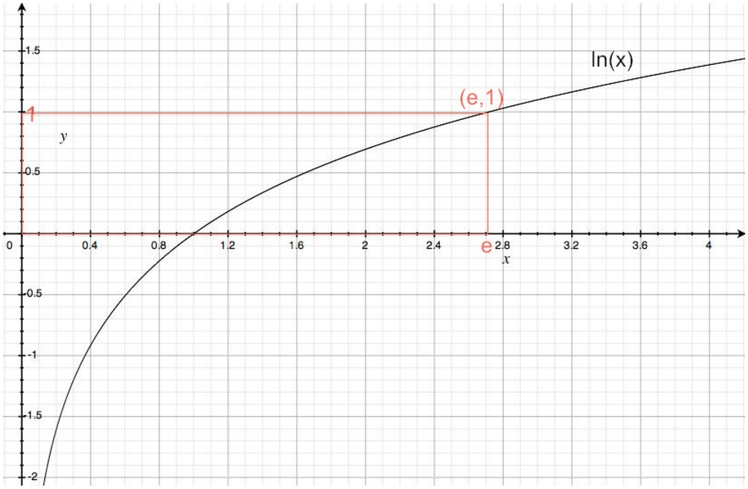


Figure 1.19: The graph of the function $\ln(x)$ passes through the following (x,y) coordinates: $(\frac{1}{e^2}, -2)$, $(\frac{1}{e}, -1)$, $(1, 0)$, $(e, 1)$, $(e^2, 2)$, $(e^3, 3)$, $(148.41\ldots, 5)$, and $(22026.46\ldots, 10)$.

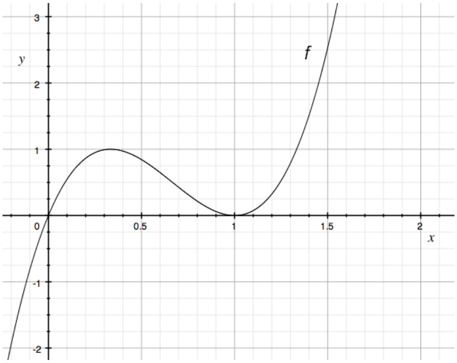
Function transformations

Often, we're asked to adjust the shape of a function by scaling it or moving it, so that it passes through certain points. For example, if we wanted to make a function g with the same shape as the absolute value function $f(x) = |x|$, but for which $g(0) = 3$, we would use the function $g(x) = |x| + 3$.

In this section, we'll discuss the four basic transformations you can perform on *any* function f to obtain a transformed function g:

- Vertical translation: $g(x) = f(x) + k$
- Horizontal translation: $g(x) = f(x - h)$
- Vertical scaling: $g(x) = Af(x)$
- Horizontal scaling: $g(x) = f(ax)$

By applying these transformations, we can *move* and *stretch* a generic function to give it any desired shape.



The next couple of pages illustrate all of the above transformations on the function

$$f(x) = 6.75(x^3 - 2x^2 + x).$$

We'll work with this function because it has distinctive features in both the horizontal and vertical directions. By observing this function's graph, we see its x-intercepts are at $x = 0$ and $x = 1$. We can confirm this mathematically by factoring the expression:

$$f(x) = 6.75x(x^2 - 2x + 1) = 6.75x(x - 1)^2.$$

The function $f(x)$ also has a local maximum at $x = \frac{1}{3}$, and the value of the function at that maximum is $f(\frac{1}{3}) = 1$.

Vertical translations

To move a function $f(x)$ *up* by k units, add k to the function:

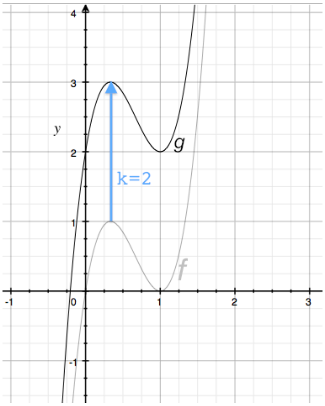
$$g(x) = f(x) + k.$$

The function $g(x)$ will have exactly the same shape as $f(x)$, but it will be *translated* (the mathematical term for moved) upward by k units.

Recall the function $f(x) = 6.75(x^3 - 2x^2 + x)$. To move the function up by $k = 2$ units, we can write

$$g(x) = f(x) + 2 = 6.75(x^3 - 2x^2 + x) + 2,$$

and the graph of $g(x)$ will be as it is shown to the right. Recall the original function $f(x)$ crosses the x-axis at $x = 0$. The transformed function $g(x)$ has the property $g(0) = 2$. The maximum at $x = \frac{1}{3}$ has similarly shifted in value from $f(\frac{1}{3}) = 1$ to $g(\frac{1}{3}) = 3$.



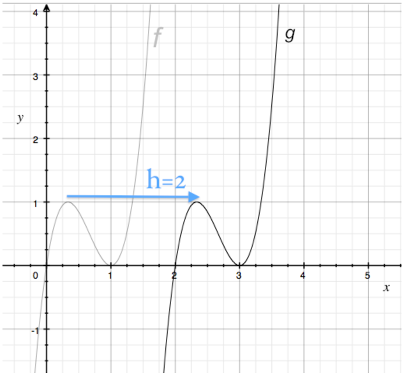
Horizontal translation

We can move a function f to the right by h units by *subtracting* h from x and using $(x - h)$ as the function's input argument:

$$g(x) = f(x - h).$$

The point $(0, f(0))$ on $f(x)$ now corresponds to the point $(h, g(h))$ on $g(x)$.

The graph to the right shows the function $f(x) = 6.75(x^3 - 2x^2 + x)$, as well as the function $g(x)$, which is shifted to the right by $h = 2$ units:



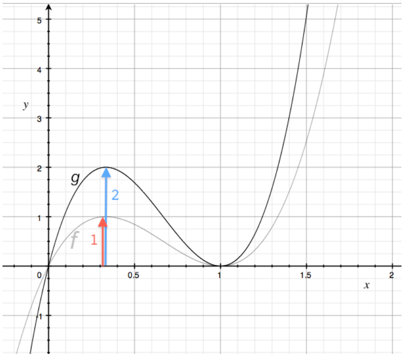
$$g(x) = f(x - 2) = 6.75\left[(x-2)^3 - 2(x-2)^2 + (x-2)\right].$$

The original function f gives us $f(0) = 0$ and $f(1) = 0$, so the new function $g(x)$ must give $g(2) = 0$ and $g(3) = 0$. The maximum at $x = \frac{1}{3}$ has similarly shifted by two units to the right, $g(2 + \frac{1}{3}) = 1$.

Vertical scaling

To stretch or compress the shape of a function vertically, we can multiply it by some constant A and obtain

$$g(x) = Af(x).$$



If $|A| > 1$, the function will be stretched. If $|A| < 1$, the function will be compressed. If A is negative, the function will flip upside down, which is a *reflection* through the x-axis.

There is an important difference between vertical translation and vertical scaling. Translation moves all points of the function by the same amount, whereas scaling moves each point proportionally to that point's distance from the x-axis.

The function $f(x) = 6.75(x^3-2x^2+x)$, when stretched vertically by a factor of $A = 2$, becomes the function

$$g(x) = 2f(x) = 13.5(x^3 - 2x^2 + x).$$

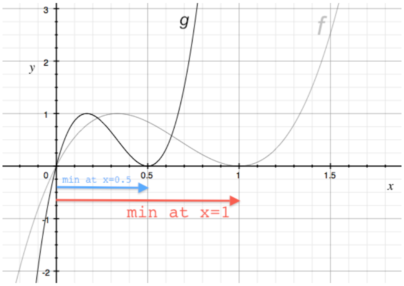
The x-intercepts $f(0) = 0$ and $f(1) = 0$ do not move, and remain at $g(0) = 0$ and $g(1) = 0$. The maximum at $x = \frac{1}{3}$ has doubled in value as $g(\frac{1}{3}) = 2$. Indeed, all values of $f(x)$ have been stretched upward by a factor of 2, as we can verify using the point $f(1.5) = 2.5$, which has become $g(1.5) = 5$.

Horizontal scaling

To stretch or compress a function horizontally, we can multiply the input value by some constant a to obtain:

$$g(x) = f(ax).$$

If $|a| > 1$, the function will be compressed. If $|a| < 1$, the function will be stretched. Note that the behaviour here is the opposite of vertical scaling. If a is a negative number, the function will also flip horizontally, which is a reflection through the y-axis.



The graph on the right shows $f(x) = 6.75(x^3 - 2x^2 + x)$, as well as the function $g(x)$, which is $f(x)$ compressed horizontally by a factor of $a = 2$:

$$\begin{aligned} g(x) &= f(2x) \\ &= 6.75\big[(2x)^3 - 2(2x)^2 + (2x)\big]. \end{aligned}$$

The x-intercept $f(0) = 0$ does not move since it is on the y-axis. The x-intercept $f(1) = 0$ does move, however, and we have $g(0.5) = 0$. The maximum at $x = \frac{1}{3}$ moves to $g(\frac{1}{6}) = 1$. All values of $f(x)$ are compressed toward the y-axis by a factor of 2.

General quadratic function

The general quadratic function takes the form

$$f(x) = A(x - h)^2 + k,$$

where x is the input, and A, h, and k are the *parameters*.

Parameters

- A: the slope multiplier

  - ▷ The larger the absolute value of A, the steeper the slope.
  - ▷ If $A < 0$ (negative), the function opens downward.
- h: the horizontal displacement of the function. Notice that subtracting a number inside the bracket $(\ )^2$ (positive h) makes the function go to the right.
- k: the vertical displacement of the function

Graph

The graph in Figure 1.20 illustrates a quadratic function with parameters $A = 1$, $h = 1$ (one unit shifted to the right), and $k = -2$ (two units shifted down).

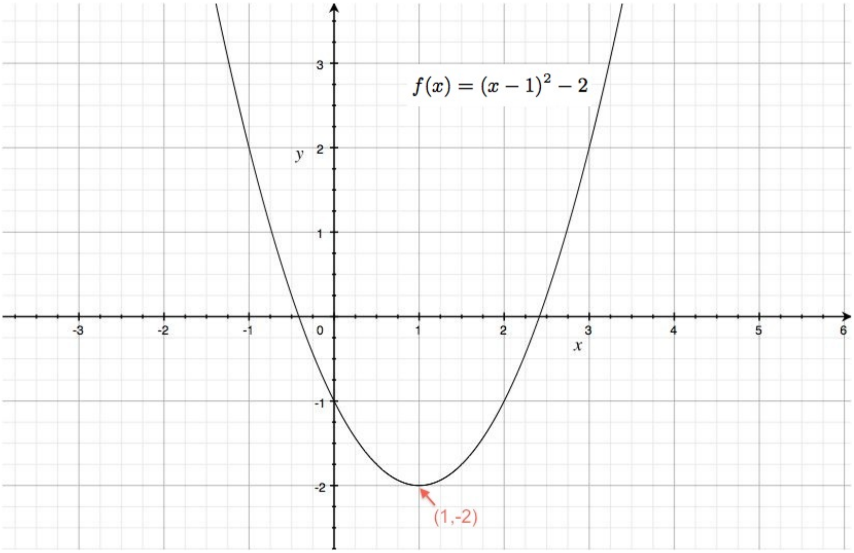


Figure 1.20: The graph of the function $f(x) = (x-1)^2 - 2$ is the same as the basic function $f(x) = x^2$, but shifted one unit to the right and two units down.

If a quadratic crosses the x-axis, it can be written in factored form:

$$f(x) = (x-a)(x-b),$$

where a and b are the two roots. Another common way of writing a quadratic function is $f(x) = Ax^2 + Bx + C$.

Properties

- There is a unique quadratic function that passes through any three points (x_1, y_1), (x_2, y_2) and (x_3, y_3), if the points have different x-coordinates: $x_1 \neq x_2$, $x_2 \neq x_3$, and $x_1 \neq x_3$.
- The derivative of $f(x) = Ax^2 + Bx + C$ is $f'(x) = 2Ax + B$.

General sine function

Introducing all possible parameters into the sine function gives us:

$$f(x) = A \sin\left(\tfrac{2\pi}{\lambda}x - \phi\right),$$

where A, λ, and ϕ are the function's parameters.

Parameters

- A: the amplitude describes the distance above and below the x-axis that the function reaches as it oscillates.
- λ: the *wavelength* of the function:

  $$\lambda \equiv \{ \text{ the distance from one peak to the next } \}.$$

- ϕ: is a phase shift, analogous to the horizontal shift h, which we have seen. This number dictates where the oscillation starts. The default sine function has zero phase shift ($\phi = 0$), so it passes through the origin with an increasing slope.

The "bare" $\sin$ function $f(x) = \sin(x)$ has wavelength 2π and produces outputs that oscillate between -1 and $+1$. When we multiply the bare function by the constant A, the oscillations will range between $-A$ and A. When the input x is scaled by the factor $\frac{2\pi}{\lambda}$, the wavelength of the function becomes λ.

1.15 Polynomials

The polynomials are a simple and useful family of functions. For example, quadratic polynomials of the form $f(x) = ax^2 + bx + c$ often arise when describing physics phenomena.

Definitions

- x: the variable
- $f(x)$: the polynomial. We sometimes denote polynomials $P(x)$ to distinguish them from generic function $f(x)$.
- Degree of $f(x)$: the largest power of x that appears in the polynomial
- Roots of $f(x)$: the values of x for which $f(x) = 0$

The most general first-degree polynomial is a line $f(x) = mx + b$, where m and b are arbitrary constants. The most general second-degree polynomial is $f(x) = a_2x^2 + a_1x + a_0$, where again a_0, a_1, and a_2 are arbitrary constants. We call a_k the *coefficient* of x^k, since this is the number that appears in front of x^k. Following the pattern, a third-degree polynomial will look like $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$.

In general, a polynomial of degree n has the equation

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0.$$

Or, if we use *summation* notation, we can write the polynomial as

$$f(x) = \sum_{k=0}^{n} a_kx^k.$$

The symbol Σ (the Greek letter *sigma*) stands for *summation*.

Solving polynomial equations

Very often in math, you will have to *solve* polynomial equations of the form

$$A(x) = B(x),$$

where $A(x)$ and $B(x)$ are both polynomials. Recall from earlier that to *solve*, we must find the value of x that makes the equality true.

Say the revenue of your company is a function of the number of products sold x, and can be expressed as $R(x) = 2x^2 + 2x$. Say also the cost you incur to produce x objects is $C(x) = x^2 + 5x + 10$. You want to determine the amount of product you need to produce to break even, that is, so that revenue equals cost: $R(x) = C(x)$. To find the break-even value x, solve the equation

$$2x^2 + 2x = x^2 + 5x + 10.$$

This may seem complicated since there are xs all over the place. No worries! We can turn the equation into its "standard form," and then use the quadratic formula. First, move all the terms to one side until only zero remains on the other side:

$$\begin{aligned}
2x^2 + 2x \quad - x^2 &= x^2 + 5x + 10 \quad - x^2 \\
x^2 + 2x \quad - 5x &= 5x + 10 \quad - 5x \\
x^2 - 3x \quad - 10 &= 10 \quad - 10 \\
x^2 - 3x - 10 &= 0.
\end{aligned}$$

Remember, if we perform the same operations on both sides of the equation, the equation remains true. Therefore, the values of x that satisfy

$$x^2 - 3x - 10 = 0,$$

namely $x = -2$ and $x = 5$, also satisfy

$$2x^2 + 2x = x^2 + 5x + 10,$$

which is the original problem we're trying to solve.

This "shuffling of terms" approach will work for any polynomial equation $A(x) = B(x)$. We can always rewrite it as $C(x) = 0$, where $C(x)$ is a new polynomial with coefficients equal to the difference of the coefficients of A and B. Don't worry about which side you move all the coefficients to because $C(x) = 0$ and $0 = -C(x)$ have exactly the same solutions. Furthermore, the degree of the polynomial C can be no greater than that of A or B.

The form $C(x) = 0$ is the *standard form* of a polynomial, and we'll explore several formulas you can use to find its solution(s).

Formulas

The formula for solving the polynomial equation $P(x) = 0$ depends on the *degree* of the polynomial in question.

First

For a first-degree polynomial,

$$P_1(x) = mx + b = 0,$$

the solution is $x = \frac{b}{m}$: just move b to the other side and divide by m.

Second

For a second-degree polynomial,

$$P_2(x) = ax^2 + bx + c = 0,$$

the solutions are $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ and $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$.

If $b^2 - 4ac < 0$, the solutions will involve taking the square root of a negative number. In those cases, we say no real solutions exist.

Higher degrees

There is also a formula for polynomials of degree 3, but it is complicated. For polynomials with order ≥ 5, there does not exist a general analytical solution.

Using a computer

When solving real-world problems, you'll often run into much more complicated equations. To find the solutions of anything more complicated than the quadratic equation, I recommend using a computer algebra system like `sympy`: `http://live.sympy.org`.

To make the computer solve the equation $x^2 - 3x + 2 = 0$ for you, type in the following:

```
>>> solve( x**2 - 3*x +2, x)          # usage: solve(expr, var)
[1, 2]
```

The function `solve` will find the roots of any equation of the form `expr = 0`. Indeed, we can verify that $x^2 - 3x + 2 = (x - 1)(x - 2)$, so $x = 1$ and $x = 2$ are the two roots.

Substitution trick

Sometimes you can solve fourth-degree polynomials by using the quadratic formula. Say you're asked to solve for x in

$$g(x) = x^4 - 3x^2 - 10 = 0.$$

Imagine this problem is on your exam, where you are not allowed the use of a computer. How does the teacher expect you to solve for x? The trick is to substitute $y = x^2$ and rewrite the same equation as

$$g(y) = y^2 - 3y - 10 = 0,$$

which you can solve by applying the quadratic formula. If you obtain the solutions $y = \alpha$ and $y = \beta$, then the solutions to the original fourth-degree polynomial are $x = \sqrt{\alpha}$ and $x = \sqrt{\beta}$, since $y = x^2$.

Since we're not taking an exam right now, we are allowed to use the computer to find the roots:

```
>>> solve(y**2 - 3*y -10, y)
[-2, 5]
>>> solve(x**4 - 3*x**2 -10 , x)
[sqrt(2)i, -sqrt(2)i, -sqrt(5) , sqrt(5) ]
```

Note how the second-degree polynomial has two roots, while the fourth-degree polynomial has four roots (two of which are imaginary, since we had to take the square root of a negative number to obtain them). The imaginary roots contain the unit imaginary number $i \equiv \sqrt{-1}$.

If you see this kind of problem on an exam, you should report the two real solutions as your answer—in this case $-\sqrt{5}$ and $\sqrt{5}$—without mentioning the imaginary solutions because you are not supposed to know about imaginary numbers yet. If you feel impatient and are ready to know about the imaginary numbers right now, feel free to skip ahead to the section on complex numbers.

1.16 Trigonometry

We can put any three lines together to make a triangle. What's more, if one of the triangle's angles is equal to $90°$, we call this triangle a *right-angle triangle*.

In this section we'll discuss right-angle triangles in great detail and get to know their properties. We'll learn some fancy new terms like *hypotenuse*, *opposite*, and *adjacent*, which are used to refer to the different sides of a triangle. We'll also use the functions *sine*, *cosine*, and *tangent* to compute the *ratios of lengths* in right triangles.

Understanding triangles and their associated trigonometric functions is of fundamental importance: you'll need this knowledge for your future understanding of mathematical subjects like vectors and complex numbers, as well as physics subjects like oscillations and waves.

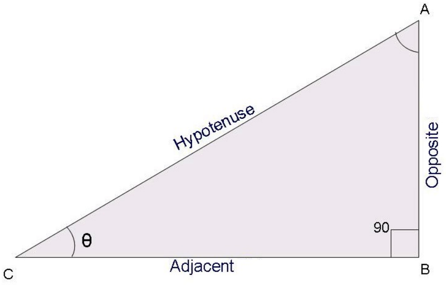


Figure 1.21: A right-angle triangle. The angle θ and the names of the sides of the triangle are indicated.

Concepts

- A, B, C: the three *vertices* of the triangle
- θ: the angle at the vertex C. Angles can be measured in degrees or radians.
- opp $\equiv \overline{AB}$: the length of the *opposite* side to θ
- adj $\equiv \overline{BC}$: the length of side *adjacent* to θ
- hyp $\equiv \overline{AC}$: the *hypotenuse*. This is the triangle's longest side.
- h: the "height" of the triangle (in this case $h =$ opp $= \overline{AB}$)
- $\sin\theta \equiv \frac{\text{opp}}{\text{hyp}}$: the *sine* of theta is the ratio of the length of the opposite side and the length of hypotenuse.
- $\cos\theta \equiv \frac{\text{adj}}{\text{hyp}}$: the *cosine* of theta is the ratio of the adjacent length and the hypotenuse length.
- $\tan\theta \equiv \frac{\sin\theta}{\cos\theta} \equiv \frac{\text{opp}}{\text{adj}}$: the *tangent* is the ratio of the opposite length divided by the adjacent length.

Pythagoras' theorem

In a right-angle triangle, the length of the hypotenuse squared is equal to the sum of the squares of the lengths of the other sides:

$$|\text{adj}|^2 + |\text{opp}|^2 = |\text{hyp}|^2.$$

If we divide both sides of the above equation by $|\text{hyp}|^2$, we obtain

$$\frac{|\text{adj}|^2}{|\text{hyp}|^2} + \frac{|\text{opp}|^2}{|\text{hyp}|^2} = 1,$$

which can be rewritten as

$$\cos^2\theta \ + \sin^2\theta = 1.$$

This is a powerful *trigonometric identity* describing the relationship between $\sin$ and $\cos$.

Sin and cos

Meet the trigonometric functions, or trigs for short. These are your new friends. Don't be shy now, say hello to them.

"Hello."

"Hi."

"Soooooo, you are like functions right?"

"Yep," sin and cos reply in chorus.

"Okay, so what do you do?"

"Who me?" asks cos. "Well I tell the ratio. . . hmm. . . Wait, are you asking what I do as a *function* or specifically what *I* do?"

"Both I guess?"

"Well, as a function, I take angles as inputs and I give ratios as answers. More specifically, I tell you how 'wide' a triangle with that angle will be," says cos all in one breath.

"What do you mean wide?" you ask.

"Oh yeah, I forgot to say, the triangle must have a hypotenuse of length 1. What happens is there is a point P that moves around on a circle of radius 1, and we *imagine* a triangle formed by the point P, the origin, and the point on the x-axis located directly below the point P."

"I am not sure I get it," you confess.

"Let me try explaining," says sin. "Look on the next page, and you'll see a circle. This is the unit circle because it has a radius of 1. You see it, yes?"

"Yes."

"This circle is really cool. Imagine a point P that starts from the point $P(0) = (1, 0)$ and moves along the circle of radius 1. The x and y-coordinates of the point $P(\theta) = (P_x(\theta),\ P_y(\theta))$ as a function of θ are

$$P(\theta) = (P_x(\theta),\ P_y(\theta)) = (\cos\theta,\ \sin\theta).$$

So, *either* you can think of us in the context of triangles, or you think of us in the context of the unit circle."

"Cool. I kind of get it. Thanks so much," you say, but in reality you are weirded out. Talking functions? "Well guys. It was nice to meet you, but I have to get going, to finish the rest of the book."

"See you later," says cos.

"Peace out," says sin.

You should be familiar with the values of $\sin$ and $\cos$ for all angles that are multiples of $\frac{\pi}{6}$ $(30°)$ or $\frac{\pi}{4}$ $(45°)$. All of them are shown in Figure 1.22. For each angle, the x-coordinate (the first number in the bracket) is $\cos$, and the y-coordinate is $\sin$.

Maybe you're thinking that's way too much to remember. Don't worry, you just have to memorize one fact:

$$\sin(30°) = \sin\!\left(\frac{\pi}{6}\right) = \frac{1}{2}\,.$$

Knowing this, you can determine all the other angles. Let's start with $\cos(30°)$. We know that at $30°$, point P on the unit circle has the vertical coordinate $\frac{1}{2} = \sin(30°)$. We also know the $\cos$ quantity we are looking for is, by definition, the horizontal component:

$$P = (\cos(30°), \sin(30°)).$$

Key fact: all points on the unit circle are a distance of 1 from the origin. Knowing that P is a point on the unit circle, and knowing the value of $\sin(30°)$, we can solve for $\cos(30°)$. Start with the following identity,

$$\cos^2\theta\ + \sin^2\theta = 1,$$

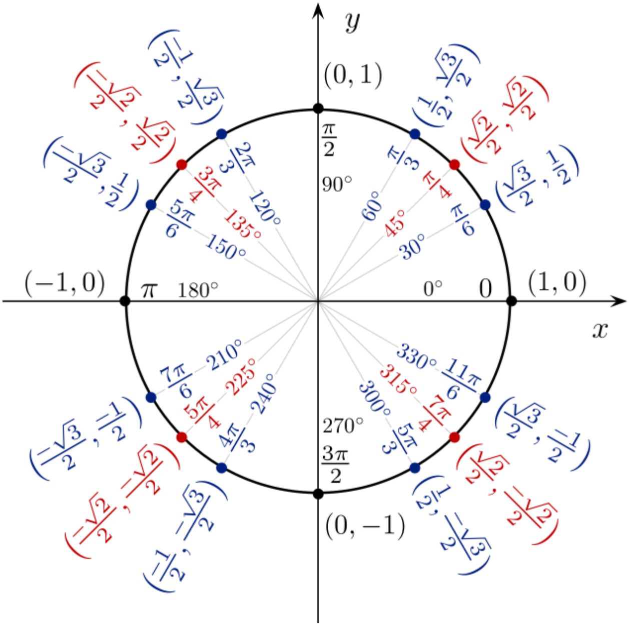

y
x
$(0,1)$
$(1,0)$
$(-1,0)$
$(0,-1)$
$\frac{\pi}{2}$
90°
π
180°
0°
0
270°
$\frac{3\pi}{2}$
$\frac{\pi}{6}$
30°
$\frac{\pi}{4}$
45°
$\frac{\pi}{3}$
60°
$\frac{2\pi}{3}$
120°
$\frac{3\pi}{4}$
135°
$\frac{5\pi}{6}$
150°
$\frac{7\pi}{6}$
210°
$\frac{5\pi}{4}$
225°
$\frac{4\pi}{3}$
240°
$\frac{5\pi}{3}$
300°
$\frac{7\pi}{4}$
315°
$\frac{11\pi}{6}$
330°
$\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(\frac{-1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(\frac{-\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(\frac{-\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(\frac{-\sqrt{3}}{2},\frac{-1}{2}\right)$
$\left(\frac{-\sqrt{2}}{2},\frac{-\sqrt{2}}{2}\right)$
$\left(\frac{-1}{2},\frac{-\sqrt{3}}{2}\right)$
$\left(\frac{1}{2},\frac{-\sqrt{3}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},\frac{-\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{3}}{2},\frac{-1}{2}\right)$

Figure 1.22: The unit circle. The coordinates of the point on the unit circle $(\cos\theta, \sin\theta)$ are indicated for several important values of the angle θ.

which is true for *all* angles θ. Moving things around, we obtain

$$\cos(30°) = \sqrt{1-\sin^2(30°)} = \sqrt{1-\frac{1}{4}} = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}\,.$$

To find the values of $\cos(60°)$ and $\sin(60°)$, observe the symmetry of the circle. 60 degrees measured from the x-axis is the same as 30 degrees measured from the y-axis. From this, we know $\cos(60°) = \sin(30°) = \frac{1}{2}$. Therefore, $\sin(60°) = \frac{\sqrt{3}}{2}$.

To find the values of sin and cos for angles that are multiples of $45°$, we need to find the value a such that

$$a^2 + a^2 = 1,$$

since at $45°$, the horizontal and vertical coordinates will be the same. Solving for a we find $a = \frac{1}{\sqrt{2}}$, but people don't like to see square roots in the denominator, so we write

$$\frac{\sqrt{2}}{2} = \cos(45^\circ) = \sin(45^\circ).$$

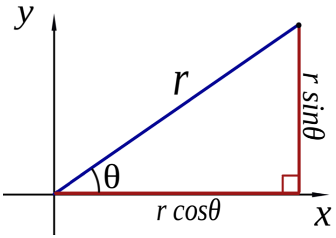
All other angles in the circle behave like the three angles above, with one difference: one or more of their components has a negative sign. For example, 150° is just like 30°, except its x component is negative. Don't memorize all the values of $\sin$ and $\cos$; if you ever need to determine their values, draw a little circle and use the symmetry of the circle to find the $\sin$ and $\cos$ components.

Non-unit circles

Consider a point $Q(\theta)$ at an angle of θ on a circle with radius $r \neq 1$. How can we find the x and y-coordinates of the point $Q(\theta)$?

We saw that the coefficients $\cos\theta$ and $\sin\theta$ correspond to the x and y-coordinates of a point on the *unit* circle ($r = 1$). To obtain the coordinates for a point on a circle of radius r, we must *scale* the coordinates by a factor of r:

$$Q(\theta) = (Q_x(\theta), Q_y(\theta)) = (r\cos\theta, r\sin\theta).$$



The take-away message is that you can use the functions $\cos\theta$ and $\sin\theta$ to find the "horizontal" and "vertical" components of any length r.

From this point on in the book, we'll always talk about the length of the *adjacent* side as $r_x = r\cos\theta$, and the length of the *opposite* side as $r_y = r\sin\theta$. It is extremely important you get comfortable with this notation.

The reasoning behind the above calculations is as follows:

$$\cos\theta \equiv \frac{\text{adj}}{\text{hyp}} = \frac{r_x}{r} \quad \Rightarrow \quad r_x = r\cos\theta,$$

´and

$$\sin\theta \equiv \frac{\text{opp}}{\text{hyp}} = \frac{r_y}{r} \quad \Rightarrow \quad r_y = r\sin\theta.$$

Calculators

Make sure to set your calculator to the correct units for working with angles. What should you type into your calculator to compute the sine of 30 degrees? If your calculator is set to degrees, simply type: $\boxed{\text{30}}$, $\boxed{\texttt{sin}}$, $\boxed{\text{=}}$.

If your calculator is set to radians, you have two options:

1. Change the `mode` of the calculator so it works in degrees.

2. Convert 30° to radians
$$30\ [^\circ] \times \frac{2\pi\ [\text{rad}]}{360\ [^\circ]} = \frac{\pi}{6}\ [\text{rad}],$$
and type: $\boxed{\pi}$, $\boxed{/}$, $\boxed{6}$, $\boxed{\texttt{sin}}$, $\boxed{\text{=}}$ on your calculator.

1.17 Trigonometric identities

There are a number of important relationships between the values of the functions $\sin$ and $\cos$. Here are three of these relationships, known as *trigonometric identities*. There about a dozen other identities that are less important, but you should memorize these three.

The three identities to remember are:

1. Unit hypotenuse

$$\sin^2(\theta) + \cos^2(\theta) = 1.$$

The unit hypotenuse identity is true by the Pythagoras theorem and the definitions of sin and cos. The ratio of the squares of the sides of a triangle are equal to the square of the size of the hypotenuse.

2. sico + sico

$$\sin(a + b) = \sin(a)\cos(b) + \sin(b)\cos(a).$$

The mnemonic for this identity is "sico + sico."

3. coco − sisi

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b).$$

The mnemonic for this identity is “coco - sisi.” The negative sign is there because it’s not good to be a sissy.

Derived formulas

If you remember the above three formulas, you can derive pretty much all the other trigonometric identities.

Double angle formulas

Starting from the sico-sico identity as explained above, and setting $a = b = x$, we can derive the following identity:

$$\sin(2x) = 2\sin(x)\cos(x).$$

Starting from the coco-sisi identity, we derive

$$\cos(2x) \;=\; 2\cos^2(x) - 1 \;= 2\left(1 - \sin^2(x)\right) - 1 = 1 - 2\sin^2(x),$$

or, if we rewrite this equation to isolate the $\sin^2$ and $\cos^2$, we get

$$\cos^2(x) = \frac{1}{2}\left(1 + \cos(2x)\right), \qquad \sin^2(x) = \frac{1}{2}\left(1 - \cos(2x)\right).$$

Self similarity

Sin and cos are periodic functions with period 2π. Adding a multiple of 2π to the function’s input does not change the function:

$$\sin(x + 2\pi) = \sin(x + 124\pi) = \sin(x), \qquad \cos(x + 2\pi) = \cos(x).$$

Furthermore, sin and cos are self similar within each 2π cycle:

$$\sin(\pi - x) = \sin(x), \qquad \cos(\pi - x) = -\cos(x).$$

Sin is cos, cos is sin

It shouldn't be surprising if I tell you that sin and cos are actually $\frac{\pi}{2}$-shifted versions of each other:

$$\cos(x) = \sin\Big(x+\frac{\pi}{2}\Big) = \sin\Big(\frac{\pi}{2}-x\Big), \quad \sin(x) = \cos\left(x-\frac{\pi}{2}\right) = \cos\left(\frac{\pi}{2}-x\right).$$

Sum formulas

$$\begin{aligned}
\sin(a) + \sin(b) &= 2\sin\left(\frac{1}{2}(a+b)\right)\cos\left(\frac{1}{2}(a-b)\right),\\
\sin(a) - \sin(b) &= 2\sin\left(\frac{1}{2}(a-b)\right)\cos\left(\frac{1}{2}(a+b)\right),\\
\cos(a) + \cos(b) &= 2\cos\left(\frac{1}{2}(a+b)\right)\cos\left(\frac{1}{2}(a-b)\right),\\
\cos(a) - \cos(b) &= -2\sin\left(\frac{1}{2}(a+b)\right)\sin\left(\frac{1}{2}(a-b)\right).
\end{aligned}$$

Product formulas

$$\begin{aligned}
\sin(a)\cos(b) &= \frac{1}{2}(\sin(a+b) + \sin(a-b)),\\
\sin(a)\sin(b) &= \frac{1}{2}(\cos(a-b) - \cos(a+b)),\\
\cos(a)\cos(b) &= \frac{1}{2}(\cos(a-b) + \cos(a+b)).
\end{aligned}$$

Discussion

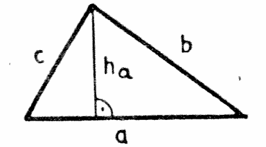
The above formulas will come in handy when you need to find some unknown in an equation, or when you are trying to simplify a trigonometric expression. I am not saying you should necessarily memorize them, but you should be aware that they exist.

1.18 Geometry

Triangles

The area of a triangle is equal to $\frac{1}{2}$ times the length of its base times its height:

$$A = \frac{1}{2} a h_a.$$

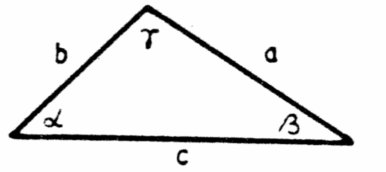


Note that h_a is the height of the triangle *relative to* the side a.

The perimeter of a triangle is

$$P = a + b + c.$$

Consider a triangle with internal angles α, β and γ. The sum of the inner angles in any triangle is equal to two right angles: $\alpha + \beta + \gamma = 180°$.



Sine rule The sine law is

$$\frac{a}{\sin(\alpha)} = \frac{b}{\sin(\beta)} = \frac{c}{\sin(\gamma)},$$

where α is the angle opposite to a, β is the angle opposite to b, and γ is the angle opposite to c.

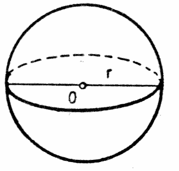
Cosine rule The cosine rules are

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc\cos(\alpha), \\ b^2 &= a^2 + c^2 - 2ac\cos(\beta), \\ c^2 &= a^2 + b^2 - 2ab\cos(\gamma). \end{aligned}$$

Sphere

A sphere is described by the equation

$$x^2 + y^2 + z^2 = r^2.$$



The surface area of a sphere is

$$A = 4\pi r^2,$$

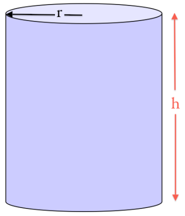
and the volume of a sphere is

$$V = \frac{4}{3}\pi r^3.$$

Cylinder

The surface area of a cylinder consists of the top and bottom circular surfaces, plus the area of the side of the cylinder:

$$A = 2\left(\pi r^2\right) + (2\pi r)h.$$



The formula for the volume of a cylinder is the product of the area of the cylinder's base times its height:

$$V = \left(\pi r^2\right) h.$$

Example You open the hood of your car and see 2.0 L written on top of the engine. The 2.0 L refers to the total volume of the four pistons, which are cylindrical in shape. The owner's manual tells you the diameter of each piston (bore) is 87.5 mm, and the height of each piston (stroke) is 83.1 mm. Verify that the total volume of the cylinder displacement of your engine is indeed 1998789 mm^3 ≈ 2 L.

Links

[Formula for calculating the distance between two points on a sphere]
http://www.movable-type.co.uk/scripts/latlong.html

1.19 Circle

The circle is a set of points located a constant distance from a centre point. This geometrical shape appears in many situations.

Definitions

- r: the radius of the circle
- A: the area of the circle
- C: the circumference of the circle
- (x, y): a point on the circle
- θ: the angle (measured from the x-axis) of some point on the circle

Formulas

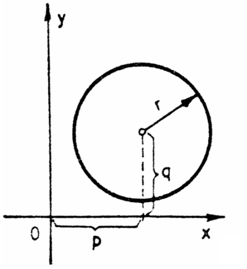
A circle with radius r centred at the origin is described by the equation

$$x^2 + y^2 = r^2.$$

All points (x, y) that satisfy this equation are part of the circle.

Rather than staying centred at the origin, the circle's centre can be located at any point (p, q) on the plane:

$$(x - p)^2 + (y - q)^2 = r^2.$$



Explicit function

The equation of a circle is a *relation* or an *implicit function* involving x and y. To obtain an *explicit function* $y = f(x)$ for the circle, we can solve for y to obtain

$$y = \sqrt{r^2 - x^2}, \quad -r \leq x \leq r,$$

and

$$y = -\sqrt{r^2 - x^2}, \quad -r \leq x \leq r.$$

The explicit expression is really two functions, because a vertical line crosses the circle in two places. The first function corresponds to the top half of the circle, and the second function corresponds to the bottom half.

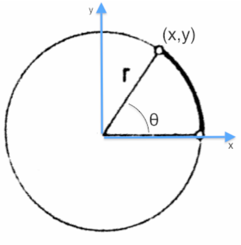
Polar coordinates

Circles are so common in mathematics that mathematicians developed a special "circular coordinate system" in order to describe them more easily.

It is possible to specify the coordinates (x, y) of any point on the circle in terms of the *polar coordinates* $r\angle\theta$, where r measures the distance of the point from the origin, and θ is the angle measured from the x-axis.

To convert from the polar coordinates $r\angle\theta$ to the (x, y) coordinates, use the trigonometric functions $\cos$ and $\sin$:

$$x = r\cos\theta \qquad \text{and} \qquad y = r\sin\theta.$$



Parametric equation

We can describe *all* the points on the circle if we specify a fixed radius r and vary the angle θ over all angles: $\theta \in [0, 360°)$. A *parametric equation* specifies the coordinates $(x(\theta), y(\theta))$ for the points on a curve, for all values of the *parameter* θ. The parametric equation for a circle of radius r is given by

$$\{(x, y) \in \mathbb{R}^2 \mid x = r\cos\theta, y = r\sin\theta,\ \theta \in [0, 360°)\}.$$

Try to visualize the curve traced by the point $(x(\theta), y(\theta)) = (r\cos\theta, r\sin\theta)$ as θ varies from $0°$ to $360°$. The point will trace out a circle of radius r.

If we let the parameter θ vary over a smaller interval, we'll obtain subsets of the circle. For example, the parametric equation for the top half of the circle is

$$\{(x, y) \in \mathbb{R}^2 \mid x = r\cos\theta, y = r\sin\theta,\ \theta \in [0, 180°]\}.$$

The top half of the circle is also described by $\{(x,y) \in \mathbb{R}^2 \mid y = \sqrt{r^2 - x^2},\ x \in [-r,r]\}$, where the parameter used is the x-coordinate.

Area

The area of a circle of radius r is $A = \pi r^2$.

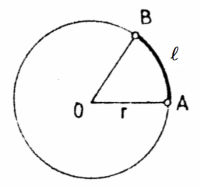
Circumference and arc length

The circumference of a circle is

$$C = 2\pi r.$$

This is the total length you can measure by following the curve all the way around to trace the outline of the entire circle.

What is the length of a part of the circle? Say you have a piece of the circle, called an *arc*, and that piece corresponds to the angle $\theta = 30°$. What is the arc's length? If the circle's total length $C = 2\pi r$ represents a full $360°$ turn around the circle, then the arc length ℓ for a portion of the circle corresponding to the angle θ is



$$\ell = 2\pi r \frac{\theta}{360}\,.$$

Note the arc length ℓ depends on the angle θ.

Radians

Though degrees are commonly used as a unit for angles, it is much better to measure angles in *radians*, since radians are the *natural* units for measuring angles. The conversion ratio from degrees to radians is 2π [radians] $= 360$ [degrees]. For a circle of radius $r = 1$, the arc length is equal to the angle in radians:

$$\ell = \theta_{\text{rad}}.$$

Measuring radians is equivalent to measuring arc length on a circle of radius 1.

1.20 Ellipse

The ellipse is a fundamental shape that occurs in nature. The orbit of planet Earth around the Sun is an ellipse.

Parameters

- a: the half-length of the ellipse along the x-axis, also known as the semi-major axis
- b: the half-length of the ellipse along the y-axis
- ε: the *eccentricity* of the ellipse, $\varepsilon \equiv \sqrt{1 - \frac{b^2}{a^2}}$
- F_1, F_2: the two *focal points* of the ellipse
- r_1: the distance from a point on the ellipse to F_1
- r_2: the distance from a point on the ellipse to F_2

Definition

An ellipse is the curve found by tracing along all the points for which the sum of the distances to the two focal points is a constant:

$$r_1 + r_2 = \text{const.}$$

There's a neat way to draw a perfect ellipse using a piece of string and two tacks or pins. Take a piece of string and tack it to a picnic table at two points, leaving some loose slack in the middle of the string. Now take a pencil, and without touching the table, use the pencil to pull the middle of the string until it is taut. Make a mark at that point. With the two parts of string completely straight, make a mark at every point possible where the two "legs" of string remain taut.

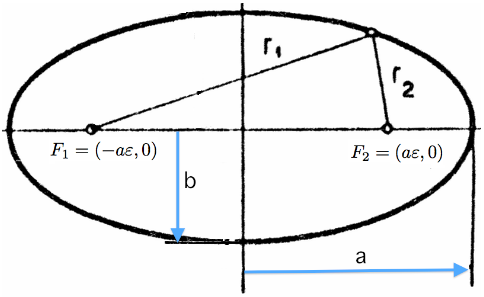


Figure 1.23: An ellipse with semi-major axis a and semi-minor axis b. The locations of the focal points F_1 and F_2 are indicated.

An ellipse is a set of points (x, y) that satisfy the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

The *eccentricity* of an ellipse describes how elongated it is:

$$\varepsilon \equiv \sqrt{1 - \frac{b^2}{a^2}}\,.$$

The parameter $\varepsilon \in [0, 1)$ describes the *shape* of the ellipse in a scale-less fashion. The bigger ε is, the bigger the difference will be between the length of the semi-major axis and the semi-minor axis. In the special case when $\varepsilon = 0$, the equation of the ellipse becomes a circle with radius a.

The (x, y)-coordinates of the two focal points are

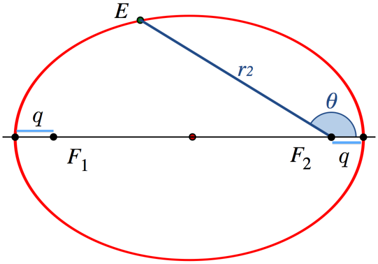
$$F_1 = (-a\varepsilon, 0) \qquad \text{and} \qquad F_2 = (a\varepsilon, 0).$$

The focal points correspond to the locations of the two tacks where the string is held in place. Recall that we defined the variables r_1 and r_2 to represent the distance from the focal points F_1 and F_2. Furthermore, we will denote by $q = a(1 - \varepsilon)$ the distance of the ellipse's closest approach to a focal point.

Polar coordinates

In polar coordinates, the ellipse can be described by a function $r_2(\theta)$. This function gives the distance of a point E from F_2 as a function of the angle θ. Recall in polar coordinates, the angle θ is the independent variable and the dependent variable is the distance $r_2(\theta)$.

The equation of the ellipse in polar coordinates depends on the length of the semi-major axis a and the eccentricity ε. The equation that describes an ellipse in polar coordinates is



$$r_2(\theta) = \frac{a(1-\varepsilon^2)}{1+\varepsilon\cos(\theta)},$$

where the angle θ is measured with respect to the positive x-axis. The distance is smallest when $\theta = 0$ with $r_2(0) = a(1-\varepsilon) = q$ and largest when $\theta = \pi$ with $r_2(\pi) = a + a\varepsilon = a(1+\varepsilon)$.

Calculating the orbit of the Earth

To a close approximation, the motion of the Earth around the Sun is described by an ellipse with the Sun positioned at the focus F_2. We can therefore use the polar coordinates formula $r_2(\theta)$ to describe the distance of the Earth from the Sun.

The eccentricity of Earth's orbit around the Sun is $\varepsilon = 0.01671123$, and the half-length of the major axis is $a = 149\,598\,261$[km]. We substitute these values into the general formula for $r_2(\theta)$ and obtain the following equation:

$$r_2(\theta) = \frac{149\,556\,484.56}{1+0.01671123\cos(\theta)} \qquad \text{[km]}.$$

The point where the Earth is closest to the Sun is called the *perihelion*. It occurs when $\theta = 0$, which happens around the 3rd of January. The moment where the Earth is most distant from the Sun is called the *aphelion* and corresponds to the angle $\theta = \pi$. Earth's *aphelion* happens around the 3rd of July.

We can use the formula for $r_2(\theta)$ to predict the *perihelion* and *aphelion* distances of Earth's orbit:

$$r_{2,\text{peri}} = r_2(0) = \frac{149556483}{1 + 0.01671123\cos(0)} = 147\,098\,290 \text{ [km]},$$
$$r_{2,\text{aphe}} = r_2(\pi) = \frac{149556483}{1 + 0.01671123\cos(\pi)} = 152\,098\,232 \text{ [km]}.$$

Can you verify that the above predictions are accurate?

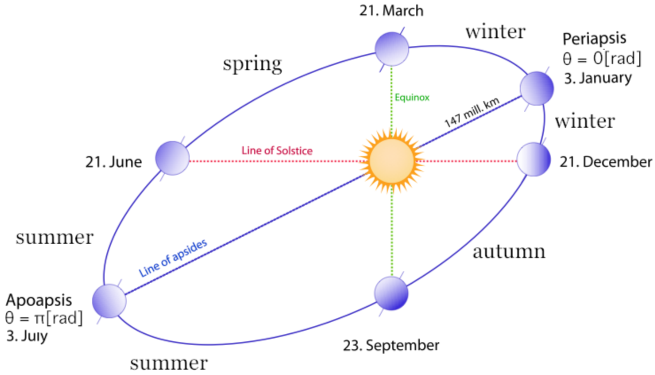


Figure 1.24: The orbit of the Earth around the Sun. Key points of the orbit are labelled. The seasons in the Northern hemisphere are also indicated.

The angle θ of the Earth relative to the Sun can be described as a function of time $\theta(t)$. The exact formula of the function $\theta(t)$ that describes the angle as a function of time is fairly complicated, so we won't go into the details. Let's simply look at some values of $\theta(t)$ with t measured in days. We'll begin on Jan 3$^{\text{rd}}$.

Newton's insight

Contrary to common belief, Newton did not discover his theory of gravitation because an apple fell on his head while sitting under a tree. What actually happened is that he started from Kepler's laws of motion, which describe the exact elliptical

t [day]	1	2	.	182	.	365	365.242199
t [date]	Jan 3	Jan 4	.	July 3	.	Jan 2	?
$\theta(t)$ [°]	0		.	180	.	359.761356	360
$\theta(t)$ [rad]	0		.	π	.	6.27902	2π

Table 1.1: The angular position of the Earth as a function of time. Note the extra amount of "day" that is roughly equal to $\frac{1}{4} = 0.25$. We account for this discrepancy by adding an extra day to the calendar once every four years.

orbit of the Earth as a function of time. Newton asked, "What kind of force would cause two bodies to spin around each other in an elliptical orbit?" He determined that the gravitational force between the Sun of mass M and the Earth of mass m must be of the form $F_g = \frac{GMm}{r^2}$. We'll discuss more about the law of gravitation in Chapter 4.

For now, let's give props to Newton for connecting the dots, and props to Johannes Kepler for studying the orbital periods, and Tycho Brahe for doing all the astronomical measurements. Above all, we owe some props to the ellipse for being such an awesome shape!

By the way, the varying distance between the Earth and the Sun is not the reason we have seasons. The ellipse had nothing to do with seasons! Seasons are predominantly caused by the *axial tilt* of the Earth. The axis of rotation of the Earth is tilted by $23.4°$ relative to the plane of its orbit around the Sun. In the Northern hemisphere, the longest day of the year is the summer solstice, which occurs around the 21th of June. On that day, the Earth's spin axis is tilted toward the Sun so the Northern hemisphere receives the most sunlight.

Links

[Further reading about Earth-Sun geometry]
http://www.physicalgeography.net/fundamentals/6h.html

1.21 Hyperbola

The hyperbola is another fundamental shape of nature. A horizontal hyperbola is the set of points (x, y) which satisfy the equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

The numbers a and b are arbitrary constants. This hyperbola passes through the points $(-a, 0)$ and $(a, 0)$. The eccentricity of this hyperbola is defined as

$$\varepsilon = \sqrt{1 + \frac{b^2}{a^2}}\,.$$

Eccentricity is an important parameter of the hyperbola, as it determines the hyperbola's shape. Recall the ellipse is also defined by an eccentricity parameter, though the formula is slightly different. This could be a coincidence—or is there a connection? Let's see.

Graph

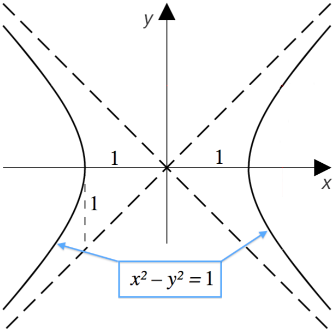


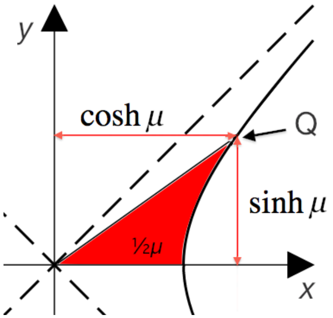
Figure 1.25: The unit hyperbola $x^2 - y^2 = 1$. The graph of the hyperbola has two branches, opening to the sides. The dashed lines are called the *asymptotes* of the hyperbola. The eccentricity determines the angle between the asymptotes. The eccentricity of $x^2 - y^2 = 1$ is $\varepsilon = \sqrt{1 + \frac{1}{1}} = \sqrt{2}$.

The graph of a hyperbola consists of two separated *branches*, as illustrated in Figure 1.25. We'll focus our discussion mostly on the right branch of the hyperbola.

Hyperbolic trigonometry

The trigonometric functions $\sin$ and $\cos$ describe the geometry of the unit circle. The point $P = (\cos\theta, \sin\theta)$ traces out the unit circle as the angle θ goes from 0 to 2π. The function $\cos$ is defined as the x-coordinate of the point P, and $\sin$ is the y-coordinate. The study of the geometry of the points on the unit circle is called *circular trigonometry*.

Instead of looking at a point P on the unit circle $x^2 + y^2 = 1$, let's trace out the path of a point Q on the unit hyperbola $x^2 - y^2 = 1$. We will now define *hyperbolic* variants of the $\sin$ and $\cos$ functions to describe the coordinates of the point Q. This is called *hyperbolic trigonometry*. Doesn't that sound awesome? Next time your friends ask what you have been up to, tell them you are learning about hyperbolic trigonometry.



The coordinates of a point Q on the unit hyperbola are $Q = (\cosh\mu, \sinh\mu)$, where μ is the *hyperbolic angle*. The x-coordinate of the point Q is $x = \cosh\mu$, and its y-coordinate is $y = \sinh\mu$. The name hyperbolic angle is a bit of a misnomer, since $\mu \in [0, \infty$ actually measures an area. The area of the highlighted region in the figure corresponds to $\frac{1}{2}\mu$.

Recall the circular-trigonometric identity $\cos^2\theta + \sin^2\theta = 1$, which follows from the fact that all the points (x, y) on the unit circle obey $x^2+y^2 = 1$. There is an analogous hyperbolic trigonometric identity:

$$\cosh^2\mu - \sinh^2\mu = 1.$$

This identity follows because we defined $x = \cosh\mu$ and $y = \sinh\mu$ to be the coordinates of a point Q which traces out the unit hyperbola $x^2 - y^2 = 1$.

The hyperbolic functions are related to the exponential function through the following formulas:

$$\cosh x = \frac{e^x + e^{-x}}{2}, \qquad \sinh x = \frac{e^x - e^{-x}}{2},$$

and

$$e^x = \cosh x + \sinh x.$$

The $\cosh$ function is even, while $\sinh$ is odd. You can think of $\cosh x$ as the "even part" of e^x, and $\sinh x$ as the "odd part" of e^x.

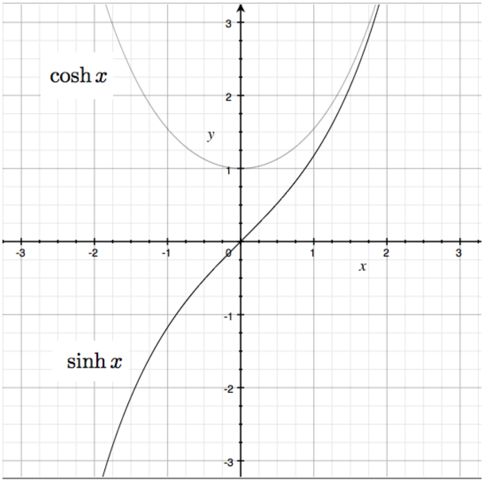


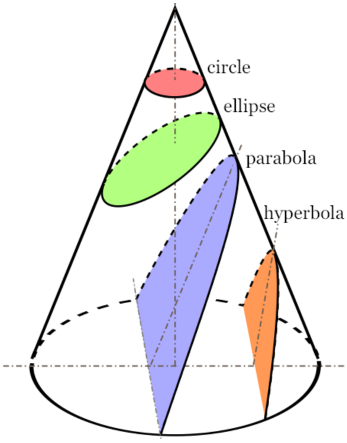
Figure 1.26: The graphs of the functions $\cosh x$ and $\sinh x$.

Don't worry about $\cosh x$ and $\sinh x$ too much. The hyperbolic trig functions are used much less often than the circular trigonometric functions $\cos\theta$ and $\sin\theta$. The main thing to remember is the general pattern: cosine functions are used to denote horizontal coordinates and sine functions are used to denote vertical coordinates.

The conic sections

There is a deep connection between the geometric shapes of the circle, the ellipse, the parabola, and the hyperbola. These seemingly different shapes can be obtained, geometrically speaking, from a single object: the cone. We can obtain the four curves by slicing the cone at different angles. Furthermore, we can use the eccentricity parameter ε to classify the curves.

A horizontal cut through the cone will produce a circle. The circle corresponds to an eccentricity parameter of $\varepsilon = 0$. For values of ε in the interval $[0, 1)$ the function $r(\theta)$ describes an ellipse. The value $\varepsilon = 1$ corresponds to the shape of a parabola. An eccentricity $\varepsilon > 1$ corresponds to the shape of a hyperbola.



Conic sections in polar coordinates

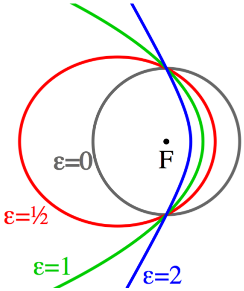
In polar coordinates, all four conic sections can be described by the same equation,

$$r(\theta) = \frac{q(1+\varepsilon)}{1+\varepsilon\cos(\theta)}\,,$$

where q is the curve's closest distance to a focal point. For a circle $q = a$, for an ellipse $q = a(1-\varepsilon)$, and for a hyperbola $q = a(\varepsilon - 1)$. In the context of a parabola, the length q is sometimes referred to as the focal length and denoted f. Depending on the parameter ε, the equation $r(\theta)$ defines either a circle, an ellipse, a parabola, or a hyperbola. Table 1.2 summarizes all our observations regarding conics.

Conic section	Equation	Polar equation	Eccentricity
Circle	$x^2 + y^2 = a^2$	$r(\theta) = a$	$\varepsilon = 0$
Ellipse	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	$r(\theta) = \frac{a(1-\varepsilon^2)}{1+\varepsilon\cos(\theta)}$	$\varepsilon = \sqrt{1 - \frac{b^2}{a^2}} \in [0,1)$
Parabola	$y^2 = 4qx$	$r(\theta) = \frac{2q}{1+\cos(\theta)}$	$\varepsilon = 1$
Hyperbola	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	$r(\theta) = \frac{a(\varepsilon^2-1)}{1+\varepsilon\cos(\theta)}$	$\varepsilon = \sqrt{1 + \frac{b^2}{a^2}} \in (1, \infty$

Table 1.2: The four conic sections and their eccentricity parameters.



The motion of the planets is explained by Newton's law of gravitation. The gravitational interaction between two bodies is always described by one of the four conic sections. The figure on the right illustrates four different trajectories for a satellite near planet F. The circle ($\varepsilon = 0$) and the ellipse ($\varepsilon \in [0,1)$) describe *closed orbits*, in which the satellite is captured in the gravitational field of the planet F and remains in orbit forever.

The parabola ($\varepsilon = 1$) and the hyperbola ($\varepsilon > 1$) describe *open orbits*, in which the satellite swings by the planet F and then continues.

Links

[Lots of great information on Wikipedia]
`http://en.wikipedia.org/wiki/Eccentricity_(mathematics)`

[An in-depth discussion on the conic sections]
`http://astrowww.phys.uvic.ca/~tatum/celmechs/celm2.pdf`

1.22 Solving systems of linear equations

We know that solving equations with one unknown—like $2x + 4 = 7x$, for instance—requires manipulating both sides of the equation until the unknown variable is *isolated* on one side. For this instance, we can subtract $2x$ from both sides of the equation to obtain $4 = 5x$, which simplifies to $x = \frac{4}{5}$.

What about the case when you are given two equations and must solve for two unknowns? For example,

$$\begin{aligned} x + 2y &= 5, \\ 3x + 9y &= 21. \end{aligned}$$

Can you find values of x and y that satisfy both equations?

Concepts

- x, y: the two unknowns in the equations
- $eq1, eq2$: a system of two equations that must be solved *simultaneously*. These equations will look like

  $$\begin{aligned} a_1x + b_1y &= c_1, \\ a_2x + b_2y &= c_2, \end{aligned}$$

  where as, bs, and cs are given constants.

Principles

If you have n equations and n unknowns, you can solve the equations simultaneously and find the values of the unknowns. There are several different approaches for solving equations simultaneously. We'll learn about three of these approaches in this section.

Solution techniques

Solving by equating

We want to solve the following system of equations:

$$\begin{aligned} x + 2y &= 5, \\ 3x + 9y &= 21. \end{aligned}$$

We can isolate x in both equations by moving all other variables and constants to the right sides of the equations:

$$\begin{aligned} x &= 5 - 2y, \\ x &= \frac{1}{3}(21 - 9y) = 7 - 3y. \end{aligned}$$

Though the variable x is still unknown, we know two facts about it: x is equal to $5 - 2y$, and x is equal to $7 - 3y$. Therefore,

$$5 - 2y = 7 - 3y.$$

We can solve for y by adding $3y$ to both sides and subtracting 5 from both sides. We find $y = 2$.

We found $y = 2$, but what is x? Easy. Plug the value $y = 2$ into any of the equations we started from. Let's try the first one:

$$x = 5 - 2y = 5 - 2(2) = 1.$$

We're done, and $x = 1, y = 2$ is our solution.

Substitution

Let's return to our set of equations to see another approach for solving:

$$\begin{aligned} x + 2y &= 5, \\ 3x + 9y &= 21. \end{aligned}$$

We can isolate x in the first equation to obtain

$$\begin{aligned} x &= 5 - 2y, \\ 3x + 9y &= 21. \end{aligned}$$

Now *substitute* the expression for x from the top equation into the bottom equation:

$$3(5 - 2y) + 9y = 21.$$

We just eliminated one of the unknowns by substitution. Continuing, we expand the bracket to find

$$15 - 6y + 9y = 21,$$

or

$$3y = 6.$$

Thus, we find $y = 2$. To solve for x, use the original equation $x = 5 - 2y$ to find $x = (5 - 2(2)) = 1$.

Subtraction

There is a third way to solve the equations

$$\begin{aligned} x + 2y &= 5, \\ 3x + 9y &= 21. \end{aligned}$$

Observe that any equation will remain true if we multiply the whole equation by some constant. For example, we can multiply the first equation by 3 to obtain an equivalent set of equations:

$$\begin{aligned} 3x + 6y &= 15, \\ 3x + 9y &= 21. \end{aligned}$$

Why did I pick 3 as the multiplier? By choosing this constant, the x terms in both equations now have the same coefficient.

Subtracting two true equations yields another true equation. Let's subtract the top equation from the bottom one:

$$3x - 3x + 9y - 6y = 21 - 15 \quad \Rightarrow \quad 3y = 6.$$

The $3x$ terms cancel. This subtraction became possible because we multiplied the first equation by 3. We see that $y = 2$. We can then substitute 2 for y in one of the original equations:

$$x + 2(2) = 5,$$

from which we deduce that $x = 1$.

Discussion

These techniques—elimination, substitution, and subtraction—can be extended to solve equations with more unknowns. There is actually an entire course called linear algebra, in which you will develop a more advanced, systematic approach for solving systems of linear equations.

1.23 Compound interest

Soon after ancient civilizations invented the notion of numbers, they started computing *interest* on loans. It is a good idea to know how interest calculations work so that you will be able to make informed decisions about your finances.

Percentages

We often talk about ratios between quantities, rather than mentioning the quantities themselves. For example, we can imagine average Joe, who invests \$1000 in the stock market and loses \$300 because the boys on Wall Street keep pulling dirty tricks on him. To put the number \$300 into perspective, we can say Joe lost 0.3 of his investment, or alternately, 30% of his investment.

To express a ratio as a percentage, multiply it by 100. The ratio of Joe's loss to investment is

$$R = 300/1000 = 0.3.$$

The same ratio expressed as a percentage gives

$$R = 300/1000 \times 100 = 30\%.$$

To convert from a percentage to a ratio, divide the percentage by 100.

Interest rates

Say you take out a $1000 loan with an interest rate of 6% compounded annually. How much will you owe in interest at the end of the year?

Since 6% corresponds to a ratio of $6/100$, and since you borrowed $1000, the accumulated interest at the end of the year will be

$$I_1 = \frac{6}{100} \times \$1000 = \$60.$$

At year's end, you'll owe the bank a total of

$$L_1 = \left(1 + \frac{6}{100}\right) 1000 = (1 + 0.06)1000 = 1.06 \times 1000 = \$1060.$$

The total money owed after 6 years will be

$$L_6 = (1.06)^6 \times 1000 = \$1418.52.$$

You borrowed $1000, but in six years you will need to give back $1418.52. This is a terrible deal! But it gets worse. The above scenario assumes that the bank compounds interest only once per year. In practice, interest is compounded each month.

Monthly compounding

An annual compounding schedule is disadvantageous to the bank, and since the bank writes the rules, compounding is usually performed every month.

The monthly interest rate can be used to find the annual rate. The bank quotes the *nominal annual percentage rate* (APR), which is equal to

$$\text{nAPR} = 12 \times r,$$

where r is the monthly interest rate.

Suppose we have a nominal APR of 6%, which gives a monthly interest rate of $r = 0.5\%$. If you borrow \$1000 at that interest rate, at the end of the first year you will owe

$$L_1 = \left(1 + \frac{0.5}{100}\right)^{12} \times 1000 = \$1061.68,$$

and after 6 years you will owe

$$L_6 = \left(1 + \frac{0.5}{100}\right)^{72} \times 1000 = 1.061677^6 \times 1000 = \$1432.04.$$

Note how the bank tries to pull a fast one: the *effective* APR is actually 6.16%, not 6%. Every twelve months, the amount due will increase by the following factor:

$$\mathrm{eAPR} = \left(1 + \frac{0.5}{100}\right)^{12} = 1.0616.$$

Thus the effective annual percent rate is $\mathrm{eAPR} = 6.16\%$.

Compounding infinitely often

What is the effective APR if the nominal APR is 6% and the bank performs the compounding n times per year?

The annual growth ratio will be

$$\left(1 + \frac{6}{100n}\right)^{n},$$

where the interest rate per compounding period is $\frac{6}{n}\%$, and there are n periods per year.

Consider a scenario in which the compounding is performed infinitely often. This corresponds to the case when the number n in the above equation tends to infinity (denoted $n \to \infty$). This is not a practical question, but it is an interesting avenue to explore nevertheless because it leads to the definition of the natural exponential function $f(x) = e^x$.

When we set $n \to \infty$ in the above expression, the annual growth ratio will be described by the exponential function base e as follows:

$$\lim_{n\to\infty}\left(1+\frac{6}{100n}\right)^n = \exp\!\left(\frac{6}{100}\right) = 1.0618365.$$

The expression "$\lim_{n\to\infty}$" is to be read as "in the limit when n tends to infinity." We will learn more about limits in Chapter 5.

A nominal APR of 6% with compounding that occurs infinitely often has an eAPR $= 6.183\%$. After six years you will owe

$$L_6 = \exp\!\left(\frac{6}{100}\right)^6 \times 1000 = \exp\!\left(\frac{36}{100}\right) \times 1000 = \$1433.33.$$

As you can see, the APR stays at a steady 6%—yet, the more frequent the compounding schedule, the more money you'll owe at the end of six years.

Links

[Very good article on interest calculations]
http://plus.maths.org/content/have-we-caught-your-interest

1.24 Set notation

A *set* is the mathematically precise notion for describing a group of objects. You need not know about sets to perform simple math; but more advanced topics require an understanding of what sets are, as well as how to denote set membership and subset relations between sets.

Definitions

- *set*: a collection of mathematical objects. The collection's contents are precisely defined.
- S, T: the usual variable names for sets
- $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$: some important sets of numbers: the naturals, the integers, the rationals, and the real numbers, respectively.
- $\{ \text{ definition } \}$: the curly brackets surround the definition of a set, and the expression inside the curly brackets describes what the set contains.

Set operations:

- $S \cup T$: the *union* of two sets. The union of S and T corresponds to the elements in either S or T.
- $S \cap T$: the *intersection* of the two sets. The intersection of S and T corresponds to the elements in both S and T.
- $S \setminus T$: *set minus*. The difference $S \setminus T$ corresponds to the elements of S that are not in T.

Set relations:

- $\subset$: is a subset of
- $\subseteq$: is a subset of or equal to

Special mathematical shorthand symbols and their corresponding meanings:

- $\forall$: for all
- $\exists$: there exists
- $\nexists$: there doesn't exist
- $|$: such that
- $\in$: element of
- $\notin$: not an element of

Sets

Much of math's power comes from *abstraction*: the ability to see the bigger picture and think *meta* thoughts about the common relationships between math objects. We can think of individual numbers like 3, -5, and π, or we can talk about the *set* of *all* numbers.

It is often useful to restrict our attention to a specific *subset* of the numbers as in the following examples.

Example 1: The nonnegative real numbers

Define $\mathbb{R}_+ \subset \mathbb{R}$ (read "$\mathbb{R}_+$ a subset of $\mathbb{R}$") to be the set of non-negative real numbers:

$$\mathbb{R}_+ \equiv \{\text{all } x \text{ in } \mathbb{R} \text{ such that } x \geq 0\},$$

or expressed more compactly,

$$\mathbb{R}_+ \equiv \{x \in \mathbb{R} \mid x \geq 0\}.$$

If we were to translate the above expression into plain English, it would read "the set $\mathbb{R}_+$ is defined as the set of all real numbers x such that x is greater or equal to zero."

Example 2: Odd and even integers

Define the set of even integers as

$$E \equiv \{n \in \mathbb{Z} \mid \tfrac{n}{2} \in \mathbb{Z}\} = \{\ldots, -2, 0, 2, 4, 6, \ldots\}$$

and the set of odd integers as

$$O \equiv \{n \in \mathbb{Z} \mid \tfrac{n+1}{2} \in \mathbb{Z}\} = \{\ldots, -3, -1, 1, 3, 5, \ldots\}.$$

In both of the above examples, we use the mathematical notation $\{\ldots \mid \ldots\}$ to define the sets. Inside the curly brackets we first describe the general kind of objects we are talking about, followed by the symbol "$|$" (read "such that"), followed by the conditions that must be satisfied by all elements of the set.

Number sets

Recall how we defined the fundamental sets of numbers in the beginning of this chapter. It is worthwhile to review them briefly.

The *natural* numbers form the set derived when you start from 0 and add 1 any number of times:

$$\mathbb{N} \equiv \{0, 1, 2, 3, 4, 5, 6, \ldots\}.$$

The integers are the numbers derived by adding or subtracting 1 some number of times:

$$\mathbb{Z} \equiv \{x \,|\, x = \pm n, n \in \mathbb{N}\}.$$

When we allow for divisions between integers, we get the rational numbers:

$$\mathbb{Q} \equiv \left\{ z \;\middle|\; z = \frac{x}{y} \text{ where } x \text{ and } y \text{ are in } \mathbb{Z}, \text{ and } y \neq 0 \right\}.$$

The broader class of real numbers also includes all rationals as well as irrational numbers like $\sqrt{2}$ and π:

$$\mathbb{R} \equiv \{\pi, e, -1.53929411\ldots,\ 4.99401940129401\ldots,\ \ldots\}.$$

Finally, we have the set of complex numbers:

$$\mathbb{C} \equiv \{1, i, 1 + i, 2 + 3i, \ldots\}.$$

Note that the definitions of $\mathbb{R}$ and $\mathbb{C}$ are not very precise. Rather than giving a precise definition of each set inside the curly brackets as we did for $\mathbb{Z}$ and $\mathbb{Q}$, we instead stated some examples of the elements in the set. Mathematicians sometimes do this and expect you to guess the general pattern for all the elements in the set.

The following inclusion relationship holds for the fundamental sets of numbers:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

This relationship means every natural number is also an integer. Every integer is a rational number. Every rational number is a real. Every real number is also a complex number.

New vocabulary

The specialized notation used by mathematicians can be difficult to get used to. You must learn how to read symbols like $\exists$, $\subset$, $|$, and $\in$ and translate their meaning in the sentence. Indeed, learning advanced mathematics notation is akin to learning a new language.

To help you practice the new vocabulary, we will look at an ancient mathematical proof and express it in terms of modern mathematical symbols.

Square-root of 2 is irrational

Claim: $\sqrt{2} \notin \mathbb{Q}$. This means there do not exist numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$. The same sentence in mathematical notation would read,

$$\nexists m, n \mid m \in \mathbb{Z}, n \in \mathbb{Z}, \text{ and } m/n = \sqrt{2}.$$

To prove this claim we will use a technique called *proof by contradiction*. We'll begin by assuming the opposite of what we want to prove: that there exist numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$. We'll then carry out some simple algebra steps and in the end we'll obtain an equation that is not true—we'll arrive at a contradiction. Arriving at a contradiction means our original supposition is wrong: there are no numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$.

Proof: Suppose there exist numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$. We can assume the integers m and n have no common factors. In particular, m and n cannot both be even, otherwise they would both contain at least one factor of 2. Next, we'll investigate whether m is an even number $m \in E$, or an odd number $m \in O$. Look back to Example 2 for the definitions of the sets O and E.

Before we check for even and oddness, it will help to point out the fact that the action of squaring an integer preserves its odd/even nature. An even number times an even number gives an even number: if $e \in E$ then $e^2 \in E$. Similarly, an odd number times an odd number gives an odd number: if $o \in O$ then $o^2 \in O$.

We proceed with the proof. We assume $m/n = \sqrt{2}$. Taking the square of both sides of this equation, we obtain

$$\frac{m^2}{n^2} = 2 \qquad \Rightarrow \qquad m^2 = 2n^2.$$

Let's analyze this last equation in more detail. If m is an odd number then m^2 will also be odd, but this would contradict the above equation since the right-hand side of the equation contains a factor of 2. Recall that any number containing a factor 2 is even, so if must be that both sides of the equation are even: $m^2 \in E$ and so $m \in E$.

If m is even then it must contain a factor of 2, so it can be written as $m = 2q$ where q is some other number $q \in \mathbb{Z}$. The exact value of q is not important. Let's revisit $m^2 = 2n^2$ once more, this time substituting $m = 2q$ into the equation:

$$(2q)^2 = 2n^2 \quad \Rightarrow \quad 2q^2 = n^2.$$

By a similar reasoning as before, we can conclude n must be an even number: $n \in E$. However, this statement contradicts our previous statement that m and n cannot both be even!

The fact that we arrived at a contradiction means we must have made a mistake somewhere in our reasoning. Since each of the steps we carried out were correct, the mistake must be in the original premise, namely that "there exist numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$." Rather, the opposite must be true: "there do not exist numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ such that $m/n = \sqrt{2}$." The last statement is equivalent to saying $\sqrt{2}$ is irrational, which is what we wanted to prove. □

Set relations and operations

Figure 1.27 illustrates the notion of a set B that is strictly contained in the set A. We say $B \subset A$ if $\forall b \in B$, we also have $b \in A$, and $\exists a \in A$ such that $a \notin B$. In other words, we write $B \subset A$ whenever the set B contains A, but there are also numbers in B that are not part of A.

Also illustrated in Figure 1.27 is the union of two sets $A \cup B$, which includes all the elements of both A and B. If $e \in A \cup B$, then $e \in A$ and/or $e \in B$.

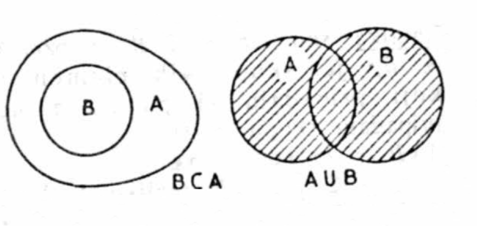


Figure 1.27: The left side of the figure is an illustration of a set B which is strictly contained in another set A, denoted $B \subset A$. The right side of the figure illustrates the union of two sets $A \cup B$.

The set intersection $A \cap B$ and set minus $A \setminus B$ are illustrated in Figure 1.28.

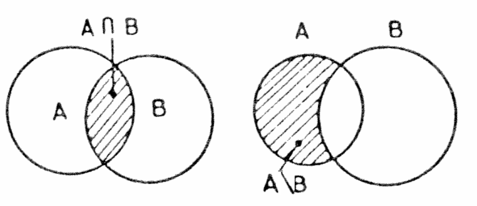


Figure 1.28: The left side of the figure shows the intersection between the sets $A \cap B$. The intersection of two sets contains the elements that are part of both sets. The right side of the figure shows the set difference $A \setminus B$, which consists of all the elements that are in A but not in B.

Sets related to functions

A function that takes real variables as inputs and produces real numbers as outputs is denoted $f : \mathbb{R} \to \mathbb{R}$. The *domain* of a function is the set of all possible inputs to the function that produce an output:

$$\mathrm{Dom}(f) \equiv \{x \in \mathbb{R} \mid f(x) \in \mathbb{R}\}.$$

Inputs for which the function is undefined are not part of the domain. For instance the function $f(x)=\sqrt{x}$ is not defined for negative inputs, so we have $\text{Dom}(f)=\mathbb{R}_+$.

The *image set* of a function is the set of all possible outputs of the function:

$$\text{Im}(f) \equiv \{y \in \mathbb{R} \mid \exists x \in \mathbb{R},\ y=f(x)\}.$$

For example, the function $f(x)=x^2$ has the image set $\text{Im}(f)=\mathbb{R}_+$ since the outputs it produces are always non-negative.

Discussion

Knowledge of the precise mathematical jargon introduced in this section is not crucial to understanding the rest of this book. That being said, I wanted to expose you to it here because this is the language in which mathematicians think. Most advanced math textbooks will assume you understand technical mathematical notation.

1.25 Math exercises

Fractions Compute the sum $1\frac{3}{4}+1\frac{31}{32}$.
Ans: $3\frac{23}{32}$

Fractions 2 Show the solution for x in the equation $\frac{1}{x}=\frac{1}{a}+\frac{1}{b}$ is $x=\frac{ab}{a+b}$.

Quadratic equation The golden ratio, denoted φ, is the positive solution to the quadratic equation $\varphi^2-\varphi-1=0$. Find the golden ratio.
Ans: $\varphi=\frac{\sqrt{5}+1}{2}$.

Swimming pool A swimming pool has length $\ell=20$ m, width $w=10$ m, and depth $d=1.5$ m. Using the fact that 1 m^3 $=1000$ L, calculate the volume of water in the swimming pool in litres?
Ans: $V=300\,000$ L.

Pythagoras' theorem Consider a right-angle triangle in which the shorter sides are 8 cm and 6 cm. What is the length of the triangle's longest side?
Ans: 10 cm.

Pythagoras' theorem 2 A television screen measures 26 inches on the diagonal. The screen height is 13 inches. How wide is the screen?
Ans: 22.51 inches.

Kepler's triangle Consider a right-angle triangle in which the hypotenuse has length $\varphi = \frac{\sqrt{5}+1}{2}$ (the golden ratio), and the adjacent side has length $\sqrt{\varphi}$. Show that the opposite side has unit length.

PDF sizing for iPad An iPad screen has a resolution of 768 pixels by 1024 pixels, and its physical dimensions are 6 inches by 8 inches. The screen has a three-to-four aspect ratio. One might conclude that the best choice of paper size for a PDF for such a screen would be 6 inches by 8 inches. At first I thought so too, but I forgot about the status bar, which is 20 pixels tall. The actual usable screen area is only 768 pixels by 1004 pixels.

Assuming the width of the PDF is chosen to be 6 inches, at what height will the PDF fit perfectly in the content area of the iPad?
Ans: The document must have a $768/1004$ aspect ratio, so its height must be $6 \times \frac{1004}{768} = 7.84375$ inches.

Formula for the quadratic equation Find the range of values of the parameter m (a real number) so that the equation $2x^2 - mx + m = 0$ has no real solutions.
Ans: The equation has no real solutions whenever $0 < m < 8$.

Chapter 2

Introduction to physics

2.1 Introduction

One of the coolest things about understanding math is that you will automatically start to understand the laws of physics too. Indeed, most physics laws are expressed as mathematical equations. If you know how to manipulate equations and you know how to solve for the unknowns in them, then you know half of physics already.

Ever since Newton figured out the whole $F = ma$ thing, people have used mechanics to achieve great technological feats, like landing spaceships on the Moon and Mars. You can be part of this science thing too. Learning physics will give you the following superpowers:

1. The power to **predict the future motion of objects** using equations. For most types of motion, it is possible to find an equation that describes the position of an object as a function of time $x(t)$. You can use this equation to predict the position of the object at all times t, including the future. "Yo G! Where's the particle going to be at $t = 1.3$ seconds?" you are asked. "It is going to be at $x(1.3)$ metres, bro." Simple as that. The equation $x(t)$ describes the object's position for *all* times t during the motion. Knowing this, you can plug $t = 1.3$ seconds into $x(t)$ to find the object's location at that time.

2. Special **physics vision** for seeing the world. After learning physics, you will start to think in term of concepts like force, acceleration, and velocity. You can use these concepts to precisely describe all aspects of the motion of objects. Without physics vision, when you throw a ball into the air you will see it go up, reach the top, then fall down. Not very exciting. Now *with* physics vision, you will see that at $t = 0$[s], the same ball is thrown in the positive y-direction with an initial velocity of $v_i = 12$[m/s]. The ball reaches a maximum height of $\max\{y(t)\} = \frac{12^2}{2\times 9.81} = 7.3$[m] at $t = 12/9.81 = 1.22$[s], then hits the ground after a total flight time of $t_f = 2\sqrt{\frac{2\times 7.3}{9.81}} = 2.44$[s]. The *measurement units* of physical quantities throughout this book are denoted in square brackets, like in the example above. Learning about the different measurement units is an important aspect of *physics vision*.

Why learn physics?

The main reason why you should learn physics is to experience *knowledge buzz*. You will learn how to calculate the motion of objects, predict the outcomes of collisions, describe oscillations, and many other useful things. As you develop your physics skills, you will be able to use physics equations to derive one physical quantity from another. For example, we can predict the maximum height reached by a ball if we know its initial velocity when thrown. Physics is a bit like playing LEGOs with a bunch of shiny new scientific building blocks.

By learning how to solve complicated physics problems, you will develop your analytical skills. Later on, you can apply these skills to other areas of life. Even if you don't go on to study science, the expertise you develop in solving physics problems will help you tackle complicated problems in general. As proof of this statement, consider the fact that companies like to hire physicists even for positions unrelated to physics: they feel confident that candidates who understand physics will be able to figure out all the business stuff easily.

Intro to science

Perhaps the most important reason you should learn physics is because it represents the golden standard for the scientific method. First of all, physics deals only with concrete things that can be **measured**. There are no feelings or subjectivities in physics. Physicists must derive mathematical models that **accurately describe** and **predict** the outcomes of experiments. Above all, we can **test** the validity of the physical models by running experiments and comparing the predicted outcome with what actually happens in the lab.

The key ingredient in scientific thinking is skepticism. Scientists must convince their peers that their equations are true without a doubt. The peers shouldn't need to *trust* the scientist; rather, they can carry out their own tests to see if the equation accurately predicts what happens in the real world. For example, let's say I *c*laim that the height of a ball thrown up in the air with speed 12[m/s] is described by the equation $y_c(t) = \frac{1}{2}(-9.81)t^2 + 12t + 0$. To test whether this equation is true, you can perform a throwing-the-ball-in-the-air experiment and record the motion of the ball as a video. You can then compare the motion parameters observed in the video with those predicted by the claimed equation $y_c(t)$.

- **Maximum height reached** One thing you can check is whether the equation $y_c(t)$ predicts the ball's maximum height $y_{\max}$. The claimed equation predicts the ball will reach its maximum height at $t = 1.22$[s]. The maximum height predicted is $\max_t\{y_c(t)\} = y_c(1.22) = 7.3$[m]. You can compare this value with the maximum height $y_{\max}$ you observe in the video.

- **Total time of flight** You can also check whether the equation $y_c(t)$ correctly predicts the time when the ball will fall back to the ground. Using the video, suppose you measure the time it took the ball to fall back to the ground to be $t_{\text{fall}} = 2.44$[s]. If the equation $y_c(t)$ is correct, it should predict a height of zero metres for the time t_{fall}.

If both predictions of the equation $y_c(t)$ match your observations from the video, you can start to believe the claimed equation of motion $y_c(t)$ is truly an accurate model for the real world.

The scientific method depends on this interplay between experiment and theory. Theoreticians prove theorems and derive equations, while experimentalists test the validity of equations. The equations that accurately predict the laws of nature are kept while inaccurate models are rejected. At the same time, experimentalists constantly measure new data and challenge theoreticians to come up with equations that correctly describe new measurements.

Equations of physics

The best physics equations are collected in textbooks. Physics textbooks contain only equations that have been extensively tested and are believed to be true. Good physics textbooks also explain how the equations are *derived* from first principles. This is important, because it is much easier to understand a few fundamental principles of physics, rather than memorize a long list of formulas. Understanding trumps memorization any day of the week.

The next section will teach you about three equations that fully describe the motion of any object: $x(t)$, $v(t)$, and $a(t)$. Using these equations and the equation-solving techniques from Chapter 1, we can predict pretty much anything we want about the position and velocity of objects undergoing *constant acceleration*.

Instead of memorizing the equations, I'll show you a cool trick for obtaining one equation of motion from another. These three equations describe different aspects of the same motion, so it's no surprise the equations are related. While you are not required to know how to derive the equations of physics, you do need to know how to use all these equations. Learning a bit of theory is a good deal: just a few pages of "difficult" theory (integrals) will give you a deep understanding of the relationship between $a(t)$, $v(t)$, and $x(t)$. This way, you can rely on your newly expanded math knowledge, rather than remember three separate formulas!

2.2 Kinematics

Kinematics (from the Greek word *kinema* for *motion*) is the study of trajectories of moving objects. The equations of kinematics can be used to calculate how long a ball thrown upward will stay in the air, or to calculate the acceleration needed to go from 0 to 100[km/h] in 5 seconds. To carry out these calculations, we need to choose the right *equation of motion* and figure out the values of the *initial conditions* (the initial position x_i and the initial velocity v_i). Afterward, we plug the known values into the appropriate equation of motion and solve for the unknown using one or two simple algebra steps. This entire section boils down to three equations and the plug-number-into-equation skill.

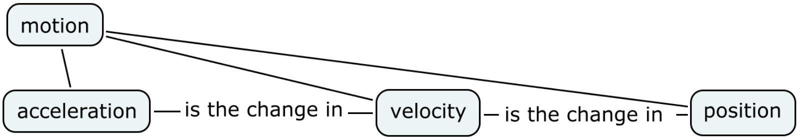


Figure 2.1: The motion of an object is described by its position, velocity, and acceleration functions.

This section is here to teach you how to use the equations of motion and help you understand the concepts of velocity and acceleration. You'll also learn how to recognize which equations to use when solving different types of physics problems.

Concepts

The key notions for describing the motion of objects are:

- t: the time. Time is measured in seconds [s].
- $x(t)$: an object's position as a function of time—also known as the equation of motion. Position is measured in metres [m] and depends on the time t.
- $v(t)$: the object's velocity as a function of time. Velocity is measured in metres per second [m/s].
- $a(t)$: the object's acceleration as a function of time. Acceleration is measured in metres per second squared [m/s^2].
- $x_i = x(0), v_i = v(0)$: the object's initial position and velocity, as measured at $t = 0$. Together x_i and v_i are known as the *initial conditions*.

Position, velocity, and acceleration

The motion of an object is characterized by three functions: the position function $x(t)$, the velocity function $v(t)$, and the acceleration function $a(t)$. The functions $x(t)$, $v(t)$, and $a(t)$ are connected—they all describe different aspects of the same motion.

You are already familiar with these notions from your experience of riding in a car. The equation of motion $x(t)$ describes the position of the car as a function of time. The velocity describes the change in the position of the car, or mathematically,

$$v(t) \equiv \text{rate of change in } x(t).$$

If we measure x in metres [m] and time t in seconds [s], then the units of $v(t)$ will be metres per second [m/s]. For example, an object moving at a constant speed of 30[m/s] will change position by 30[m] each second.

The rate of change of an object's velocity is called *acceleration*:

$$a(t) \equiv \text{rate of change in } v(t).$$

Acceleration is measured in metres per second squared [m/s^2]. A constant positive acceleration means the velocity of the motion is steadily increasing, similar to pressing the gas pedal. A constant negative acceleration means the velocity is steadily decreasing, similar to pressing the brake pedal.

In a couple of paragraphs, we'll discuss the exact mathematical equations for $x(t)$, $v(t)$, and $a(t)$, but before we dig into the math, let's look at the example of the motion of a car illustrated in Figure 2.2.

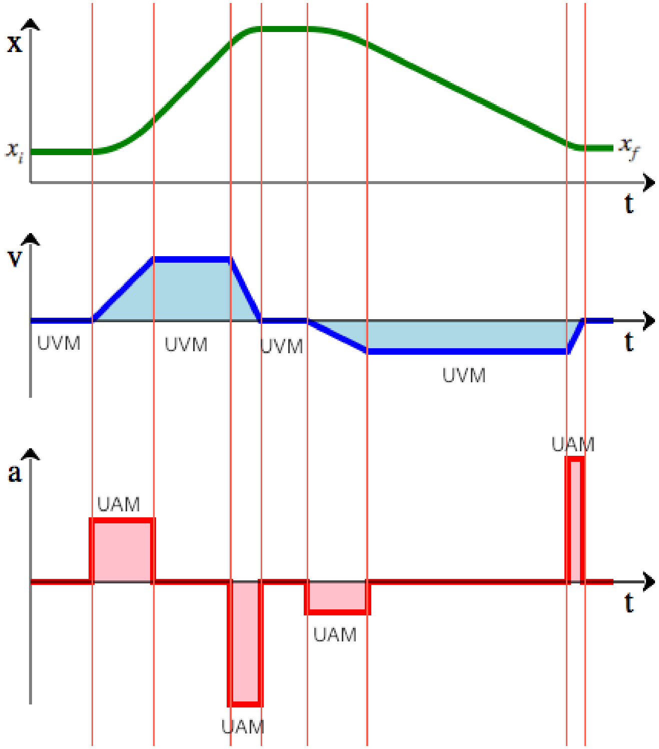


Figure 2.2: The illustration shows the simultaneous graphs of the position, velocity, and acceleration of a car during some time interval. The car starts from an initial position x_i where it sits still for some time. The driver then floors the pedal to produce a maximum acceleration for some time, and the car picks up speed. The driver then releases the accelerator, keeping it pressed enough to maintain a constant speed. Suddenly the driver sees a police vehicle in the distance and slams on the brakes (negative acceleration) and shortly afterward brings the car to a stop. The driver waits for a few seconds to make sure the cops have passed. Next, the driver switches into reverse gear and adds gas. The car accelerates backward for a bit, then maintains a constant backward speed for an extended period of time. Note how "moving backward" corresponds to negative velocity. In the end the driver slams on the brakes again to stop the car. Notice that braking corresponds to positive acceleration when the motion is in the negative direction. The car's final position is x_f.

We can observe two distinct types of motion in the situation described in Figure 2.2. During some times, the car undergoes motion at a constant velocity (uniform velocity motion, UVM). During other times, the car undergoes motions with constant acceleration (uniform acceleration motion, UAM). There exist many other types of motion, but for the purpose of this section we'll focus on these two types of motion.

- UVM: During times when there is no acceleration, the car maintains a uniform velocity and therefore $v(t)$ is a constant function. For motion with constant velocity, the position function is a line with a constant slope because, by definition, $v(t) =$ slope of $x(t)$.
- UAM: During times where the car experiences a constant acceleration $a(t) = a$, the velocity of the function changes at a constant rate. The rate of change of the velocity is constant $a =$ slope of $v(t)$, so the velocity function looks like a line with slope a. The position function $x(t)$ has a curved shape (quadratic) during moments of constant acceleration.

Formulas

There are basically four equations you need to know for this entire section. Together, these four equations fully describe all aspects of motion with constant acceleration.

Uniformly accelerated motion (UAM)

If the object undergoes a *constant* acceleration $a(t) = a$—like a car when you floor the *accelerator*—then its motion can be described by the following equations:

$$a(t) = a, \tag{2.1}$$

$$v(t) = at + v_i, \tag{2.2}$$

$$x(t) = \tfrac{1}{2}at^2 + v_i t + x_i, \tag{2.3}$$

where v_i is the initial velocity of the object and x_i is its initial position.

Here is another useful equation to remember:

$$[v(t)]^2 = v_i^2 + 2a[x(t) - x_i],$$

which is usually written

$$v_f^2 = v_i^2 + 2a\Delta x, \tag{2.4}$$

where v_f denotes the final velocity (at $t = t_f$) and Δx denotes the *change* in the x-coordinate between $t = 0$ and $t = t_f$. The triangle thing Δ is the capital Greek letter *delta*, which is often used to denote the change in quantities. Using this notation, we can rewrite Formula (2.2) in Δ-notation as $\Delta v = a\Delta t$, where $\Delta v \equiv v_f - v_i$ and $\Delta t \equiv t_f - t_i$.

That is it. Memorize these equations, plug-in the right numbers, and you can solve any kinematics problem humanly imaginable.

Uniform velocity motion (UVM)

The special case where there is zero acceleration ($a = 0$), is called *uniform velocity motion* or UVM. The velocity stays uniform (constant) because there is no acceleration. The following three equations describe the motion of an object with uniform velocity:

$$\begin{aligned} a(t) &= 0, \\ v(t) &= v_i, \\ x(t) &= v_i t + x_i. \end{aligned}$$

As you can see, these are really the same equations as in the UAM case above, but because $a = 0$, some terms are missing.

Free fall

We say an object is in *free fall* if the only force acting on it is the force of gravity. On the surface of the Earth, the force of gravity produces a constant acceleration of $a_y = -9.81$[m/s^2]. The negative sign is there because the gravitational acceleration is directed downward, and we assume the y-axis points upward. Since the gravitational acceleration is constant, we can use the UAM equations to find the height $y(t)$ and velocity $v(t)$ of objects in free fall.

Examples

Now we'll illustrate how the equations of kinematics are used.

Moroccan example Suppose your friend wants to send you a ball wrapped in aluminum foil by dropping it from his balcony, which is located at a height of $y_i = 44.145$[m]. How long will it take for the ball to hit the ground?

We recognize this is a problem with acceleration, so we start by writing the general UAM equations:

$$\begin{aligned} y(t) &= \tfrac{1}{2}at^2 + v_i t + y_i, \\ v(t) &= at + v_i. \end{aligned}$$

To find the answer, substitute the following known values into the $y(t)$ equation: $y(0) = y_i = 44.145$[m]; $a = -9.81$ (since the ball is in free fall); and $v_i = 0$[m/s] (since the ball was released from rest). We want to find the time t_{fall} when the height of the ball will be zero:

$$\begin{aligned} 0 &= y(t_{\text{fall}}), \\ 0 &= \tfrac{1}{2}(-9.81)(t_{\text{fall}})^2 + 0(t_{\text{fall}}) + 44.145. \end{aligned}$$

Solving for t_{fall} we find the answer $t_{\text{fall}} = \sqrt{\frac{44.145 \times 2}{9.81}} = 3$[s].

As another variation of this type of kinematics question, suppose you're given the time it takes for the ball to fall $t_{\text{fall}} = 3$[s], and you're asked to find the height of the balcony. You already know $y(3) = 0$, and are looking for the initial height y_i. You can solve for y_i in the equation $0 = \frac{1}{2}(-9.81)3^2 + y_i$. The answer gives $y_i = 44.145$[m].

0 to 100 in 5 seconds Say you're in the driver's seat of a car and you want to accelerate from 0 to 100[km/h] in 5 seconds. How much acceleration must the car's engine produce, assuming it produces a constant amount of acceleration?

We can calculate the necessary a by plugging the required values into the velocity equation for UAM:

$$v(t) = at + v_i.$$

Before we tackle that, we need to convert the velocity in [km/h] to velocity in [m/s]: 100[km/h] $= \frac{100[\cancel{\text{km}}]}{1[\cancel{\text{h}}]} \cdot \frac{1000[\text{m}]}{1[\cancel{\text{km}}]} \cdot \frac{1[\cancel{\text{h}}]}{3600[\text{s}]}$= 27.8 [m/s]. We substitute the desired values $v_f = 27.8$[m/s], $v_i = 0$, and $t = 5$[s] into the equation for $v(t)$ and solve for a:

$$27.8 = v(5) = a5 + 0.$$

After solving for a, we find the car's engine must produce a constant acceleration of $a = \frac{27.8}{5} = 5.56$[m/s^2] or greater.

Moroccan example II Some time later, your friend wants to send you another aluminum ball from his apartment located on the 14th floor (height of 44.145[m]). To decrease the time of flight, he *throws* the ball straight down with an initial velocity of 10[m/s]. How long does it take for the ball to hit the ground?

Imagine the apartment building as a y-axis that measures distance upward starting from the ground floor. We know the balcony is located at a height of $y_i = 44.145$[m], and that at $t = 0$[s] the ball starts with $v_i = -10$[m/s]. The initial velocity is negative because it points in the opposite direction of the y-axis. We also know there is an acceleration due to gravity of $a_y = -9.81$[m/s^2].

We start by writing the general UAM equation:

$$y(t) = \tfrac{1}{2}a_y t^2 + v_i t + y_i.$$

To find the time when the ball will hit the ground, we must solve for t in the equation $y(t) = 0$. Plug all the known values into the UAM equation,

$$y(t) = 0 = \tfrac{1}{2}(-9.81)t^2 - 10t + 44.145,$$

and solve for t using the quadratic formula. First, rewrite the quadratic equation in standard form:

$$0 = \underbrace{4.905}_{a} t^2 + \underbrace{10.0}_{b}\ t - \underbrace{44.145}_{c}.$$

Then solve using the quadratic equation:

$$t_{\text{fall}} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-10 \pm \sqrt{25 + 866.12}}{9.81} = 2.53 \qquad \text{[s]}.$$

We ignore the negative-time solution because it corresponds to a time in the past. Compared to the first Moroccan example, we see that throwing the ball downward makes it fall to the ground faster.

Discussion

Most kinematics problems you'll be asked to solve will follow the same pattern as the examples above. Given some initial values, you will be asked to solve for some unknown quantity.

It is important to keep in mind the *signs* of the numbers you plug into the equations. You should always draw the coordinate system and indicate clearly (to yourself) the x-axis, which measures the object's displacement. A velocity or acceleration quantity that points in the same direction as the x-axis is a positive number, while quantities pointing in the opposite direction are negative numbers.

By the way, all this talk about $v(t)$ being the "rate of change of $x(t)$" is starting to get on my nerves. The expression "rate of change of" is an indirect way of saying the calculus term *derivative*. In order to use this more precise terminology throughout the remainder of the book, we will now take a short excursion into the land of calculus to define two fundamental concepts: derivatives and integrals.

2.3 Introduction to calculus

Calculus is the study of functions and their properties. The two operations in the study of calculus are derivatives—which describe how quantities *change over time*—and integrals, which are used to calculate the total amount of a quantity *accumulated* over a time period.

Derivatives

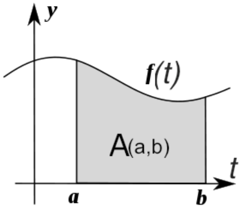
The derivative function $f'(t)$ describes how the function $f(t)$ changes over time. The derivative encodes the information about the *rate of change*, which is the *slope* of the function $f(t)$:

$$f'(t) \equiv \text{slope}_f(t) = \frac{\text{change in } f(t)}{\text{change in } t} = \frac{f(t+\Delta t) - f(t)}{\Delta t}.$$

If the derivative $f'(t)$ is equal to 5 units per second, this means that $f(t)$ changes by 5 units each second.

The derivative operation is denoted by several names and symbols: $Df(t) = f'(t) = \frac{df}{dt} = \frac{d}{dt}\{f(t)\} = \dot{f}$, all of which carry the same meaning. Think of $f'(t)$ not as a separate entity from $f(t)$, but as a *property* of the function $f(t)$. It's best to think of the derivative as an operator $\frac{d}{dt}$ that you can apply to any function to obtain its slope information. Derivatives are used in many areas of science.

Integrals



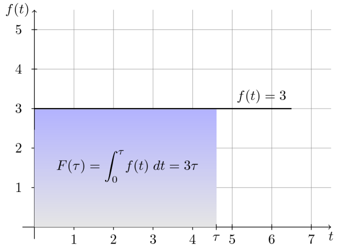
An integral corresponds to the computation of the *area* enclosed between the curve $f(t)$ and the x-axis over some interval:

$$A(a,b) \equiv \int_{t=a}^{t=b} f(t)\; dt.$$

The symbol $\int$ is shorthand for *sum*. Indeed the area under the curve corresponds to the sum of the values of the function $f(t)$ between $t=a$ and $t=b$. The integral is the total of f between a and b.

Example 1

We can easily find the area under the constant function $f(t) = 3$ between any two points because the region under the curve is rectangular. We choose to use $t = 0$ as the reference point and compute the integral $F(\tau)$, which corresponds to the area under $f(t)$ starting from $t = 0$ and going until $t = \tau$:



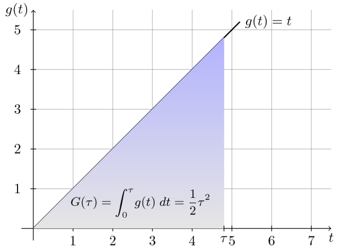
$$F(\tau) \equiv A(0, \tau) = \int_0^\tau f(t)\;dt = 3\tau.$$

The area is equal to the rectangle's height times its width.

Example 2

Consider the area under the line $g(t) = t$. Since the region under the curve is triangular, we can compute its area. Recall the area of a triangle is given by the length of its base times its height divided by 2.

We choose $t = 0$ as our starting point again and find the general formula for the area below $g(t)$ from $t = 0$ until $t = \tau$:



$$G(\tau) \equiv A(0, \tau) = \int_0^\tau g(t)\;dt = \frac{\tau \times \tau}{2} = \tfrac{1}{2}\tau^2.$$

We are able to compute the above integrals thanks to the simple geometry of the areas under the curves. Later in this book, we'll develop techniques for finding integrals of more complicated functions. In fact, there is an entire course called Integral Calculus, which is dedicated to the task of finding integrals.

What I need you to remember for now is that the integral of a function gives you the area under the curve, which is in some sense the total amount of the function accumulated during that period.

You should also remember the following two formulas:

$$\int_0^\tau a\,dt = a\tau \qquad \text{and} \qquad \int_0^\tau at\,dt = \tfrac{1}{2}a\tau^2.$$

The second formula is a generalization of the formula we saw in Example 2.

Using the above formulas in combination, you can now compute the integral under an arbitrary line $h(t) = mt + b$ as follows:

$$H(\tau) = \int_0^\tau h(t)\,dt = \int_0^\tau (mt+b)\,dt = \int_0^\tau mt\,dt \; + \int_0^\tau b\,dt = \tfrac{1}{2}m\tau^2 + b\tau.$$

Do we really need integrals? How often will you need to compute the area below a function $f(t)$ in the real world? It turns out that "calculating the area under a curve" is a very useful operation because it allows us to undo the derivative operation. Understanding the relationship between the derivative and the integral operations will allow you to solve many problems.

Inverse operations

The integral is the inverse operation of the derivative. You should already be familiar with the inverse relationship between functions. When solving equations, we use inverse functions to *undo* functions that stand in our way as we try to isolate the unknown x. Similarly, we use the integral operation to *undo* the effects of the derivative operation when we try to solve for some unknown function $f(t)$. For example, suppose $g(t)$ is a known function and we're trying to solve for $f(t)$ in the equation

$$\frac{d}{dt}\left\{f(t)\right\} = g(t).$$

Taking the integral on the left-hand side of the equation will undo the derivative operation. To keep the equality true, we must apply the integration operation on both sides of the equation to obtain

$$\begin{aligned}\int \frac{d}{dt}\{f(t)\}\,dt &= \int g(t)\,dt,\\ f(t) &= \int g(t)\,dt.\end{aligned}$$

Every time you want to *undo* a derivative, you can apply the integral operation.

There is a little technical complication we must discuss. The integral isn't *exactly* the inverse of the derivative—there exists a tricky dependence on the limits of integration.

Let's analyze in more detail what happens when we perform the combo of the derivative operation followed by the integral operation on some function $f(t)$. Suppose we are given the derivative function $f'(t)$ and asked to integrate it between $t = 0$ and $t = \tau$. Intuitively, this integral corresponds to calculating the **total of the changes** in $f(t)$ during that time interval. Recall the notation for "change in f" $\Delta f \equiv f(\tau) - f(0)$, which we used previously. This notation makes it easy to see how the integral over $f'(t)$ corresponds to the total change in $f(t)$ between $t = 0$ and $t = \tau$:

$$\int_0^\tau f'(t)\,dt = \Delta f = f(\tau) - f(0).$$

Rewriting this equation to isolate $f(\tau)$ we obtain

$$f(\tau) = \int_0^\tau f'(t)\,dt \quad + f(0).$$

The answer depends on the value of $f(t)$ at $t = 0$, the initial conditions.

Banking example To illustrate how derivative and integral operations apply to the real world, I'll draw an analogy from a scenario that every student is familiar with. Consider the function $\text{ba}(t)$, which represents your bank account balance at time t. Also consider the function $\text{tr}(t)$, which corresponds to the transactions (deposits and withdrawals) on your account.

The function $\mathrm{tr}(t)$ is the derivative of the function $\mathrm{ba}(t)$. If you ask, "how does my balance change over time?" the answer is the function $\mathrm{tr}(t)$. Using mathematical symbols, we can represent this relationship as

$$\mathrm{tr}(t) = \frac{d}{dt}\left\{\mathrm{ba}(t)\right\}.$$

If the derivative is positive, your account balance is growing. If the derivative is negative, your account balance is depleting.

Suppose you have a record of all the transactions on your account $\mathrm{tr}(t)$, and you want to compute the final account balance at the end of the month. Since $\mathrm{tr}(t)$ is the derivative of $\mathrm{ba}(t)$, you can use an integral (the inverse operation of the derivative) to obtain $\mathrm{ba}(t)$. Knowing the balance of your account at the beginning of the month, you can predict the balance at the end of the month by calculating the following integral:

$$\mathrm{ba}(30) = \mathrm{ba}(0) + \int_0^{30} \mathrm{tr}(t)\,dt.$$

This calculation makes sense since $\mathrm{tr}(t)$ represents the instantaneous changes in $\mathrm{ba}(t)$. If you want to find the overall change from day 0 until day 30, you can compute the total of all the changes in the account balance.

We use integrals every time we need to calculate the total of some quantity over a time period. In the next section, we'll see how these integration techniques can be applied to the subject of kinematics, and how the equations of motion for UAM are derived from first principles.

2.4 Kinematics with calculus

To carry out kinematics calculations, all we need to do is plug the initial conditions (x_i and v_i) into the correct equation of motion. But how did Newton come up with these equations in the first place? Now that you know Newton's mathematical techniques (calculus), you can see for yourself how the equations of motion are derived.

Concepts

Recall the kinematics concepts related to the motion of objects:

- t: time
- $x(t)$: position as a function of time
- $v(t)$: velocity as a function of time
- $a(t)$: acceleration as a function of time
- $x_i = x(0), v_i = v(0)$: the initial conditions

Position, velocity, and acceleration revisited

Recall that equations of kinematics are used to predict the motion of objects. Suppose you know the acceleration of the object $a(t)$ at all times t. Can you find $x(t)$ starting from $a(t)$?

The equations of motion $x(t)$, $v(t)$, and $a(t)$ are related:

$$a(t) \xleftarrow{\frac{d}{dt}} v(t) \xleftarrow{\frac{d}{dt}} x(t).$$

The velocity function is the derivative of the position function and the acceleration function is the derivative of the velocity function.

General procedure

If you know the acceleration of an object as a function of time $a(t)$, and you know its initial velocity $v_i = v(0)$, you can find its velocity function $v(t)$ for all later times. This is because the acceleration function $a(t)$ describes the change in the object's velocity. If you know the object started with an initial velocity of $v_i \equiv v(0)$, the velocity at a later time $t = \tau$ is equal to v_i plus the total acceleration of the object between $t = 0$ and $t = \tau$:

$$v(\tau) = v_i + \int_0^{\tau} a(t)\, dt.$$

If you know the initial position x_i and the velocity function $v(t)$, you can find the position function $x(t)$ by using integration. We find the position at time $t = \tau$ by adding all the velocities (the changes in the object's position) that occurred between $t = 0$ and $t = \tau$:

$$x(\tau) = x_i + \int_0^\tau v(t)\,dt.$$

The procedure for finding $x(t)$ starting from $a(t)$ can be summarized as follows:

$$a(t) \quad \overset{v_i + \int dt}{\longrightarrow} \quad v(t) \quad \overset{x_i + \int dt}{\longrightarrow} \quad x(t).$$

Next, I'll illustrate how you can apply this procedure to the important special case of an object undergoing uniformly accelerated motion.

Derivation of the UAM equations of motion

Consider an object undergoing uniformly accelerated motion (UAM) with acceleration function $a(t) = a$. Suppose we know the initial velocity of $v_i \equiv v(0)$, and we want to find the velocity at a later time $t = \tau$. We compute the following integral:

$$v(\tau) = v_i + \int_0^\tau a(t)\ dt = v_i + \int_0^\tau a\ dt = v_i + a\tau.$$

Velocity as a function of time is given by the initial velocity v_i added to the integral of the acceleration. The integration step can be visualized as the calculation of the area of a rectangle.

You can also use integration to find the position function $x(t)$ if you know the initial position x_i and the velocity function $v(t)$. The formula is

$$x(\tau) = x_i + \int_0^\tau v(t)\,dt = x_i + \int_0^\tau (at + v_i)\ dt = x_i + \tfrac{1}{2}a\tau^2 + v_i\tau.$$

The integration step can be visualized as the calculation of the area of a triangle with slope a stacked on top of a rectangle of height v_i.

Note that the above calculations required knowing the initial conditions x_i and v_i. These initial values were required because the integral calculations we performed only told us the *change* in the quantities relative to their initial values.

The fourth equation

We can derive the fourth equation of motion,

$$v_f^2 = v_i^2 + 2a(x_f - x_i),$$

by combining the equations of motion $v(t)$ and $x(t)$. Let's see how. Start by squaring both sides of the velocity equation $v_f = v_i + at$ to obtain

$$v_f^2 = (v_i + at)^2 = v_i^2 + 2av_it + a^2t^2 = v_i^2 + 2a[v_it + \tfrac{1}{2}at^2].$$

The term in the square bracket is equal to $\Delta x = x(t) - x_i = x_f - x_i$.

Discussion

According to Newton's second law of motion, forces are the cause of acceleration and the formula that governs this relationship is

$$F_{\text{net}} = ma,$$

where F_{net} is the magnitude of the net force acting on the object.

In Chapter 4 we'll learn about *dynamics*, the study of the different kinds of forces that can act on objects: gravitational force $\vec{F}_g$, spring force $\vec{F}_s$, friction force $\vec{F}_f$, and other forces. To find an object's acceleration, we must add together all the forces acting on the object and divide by the object's mass:

$$\sum F_i = F_{\text{net}}, \qquad \Rightarrow \qquad a = \frac{1}{m}F_{\text{net}}.$$

The physics procedure for predicting the motion of objects can be summarized as follows:

$$\frac{1}{m}\underbrace{\left(\sum \vec{F} = \vec{F}_{\text{net}}\right)}_{\text{dynamics}} = \underbrace{a(t) \overset{v_i+\int dt}{\longrightarrow} v(t) \overset{x_i+\int dt}{\longrightarrow} x(t)}_{\text{kinematics}}.$$

Free fall revisited

The force of gravity acting on an object of mass m on the surface of the Earth is given by $\vec{F}_g = -mg\hat{y}$, where $g = 9.81$[m/s^2] is the *gravitational acceleration* on the surface of the Earth. We previously discussed that an object is in *free fall* when the only force acting on it is the force of gravity. In this case, Newton's second law tells us

$$\begin{aligned} \vec{F}_{\text{net}} &= m\vec{a} \\ -mg\hat{y} &= m\vec{a}. \end{aligned}$$

Dividing both sides by the mass, we see the acceleration of an object in free fall is $\vec{a} = -9.81\hat{y}$.

It's interesting to note that an object's mass does not affect its acceleration during free fall. The force of gravity is proportional to the mass of the object, but acceleration is inversely proportional to the mass of the object; overall, it holds that $a_y = -g$ for objects in free fall, regardless of their mass. This observation was first made by Galileo in his famous Leaning Tower of Pisa experiment. Galileo dropped a wooden ball and a metal ball (same shape, different mass) from the Leaning Tower of Pisa, and observed that they fell to the ground at the same time. Search for "Apollo 15 feather and hammer drop" on `YouTube` to see this experiment performed on the Moon.

What next?

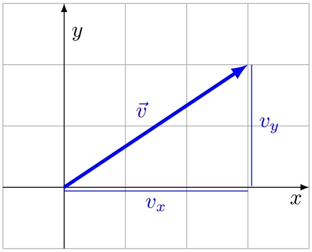
You might have noticed that in the last couple of paragraphs we started putting little arrows on top of certain quantities. The arrows are there to remind you forces, velocities, and accelerations are *vector quantities*. Before we proceed with the physics lessons, we'll make an interesting mathematical digression to introduce vectors.

Chapter 3

Vectors

In this chapter we will learn how to manipulate multi-dimensional objects called vectors. Vectors are the precise way to describe directions in space. We need vectors in order to describe physical quantities like the velocity of an object, its acceleration, and the net force acting on the object.

Vectors are built from ordinary numbers, which form the *components* of the vector. You can think of a vector as a list of numbers, and *vector algebra* as operations performed on the numbers in the list. Vectors can also be manipulated as geometrical objects, represented by arrows in space. The arrow that corresponds to the vector $\vec{v} = (v_x, v_y)$ starts at the origin $(0,0)$ and ends at the point (v_x, v_y). The word vector comes from the Latin *vehere*, which means *to carry*. Indeed, the vector $\vec{v}$ takes the point $(0,0)$ and carries it to the point (v_x, v_y).



This chapter will introduce you to vectors, vector algebra, and vector operations, which are very useful for solving physics problems. What you'll learn here applies more broadly to problems in computer graphics, probability theory, machine learning, and other fields of science and mathematics. It's all about vectors these days, so you better get to know them.

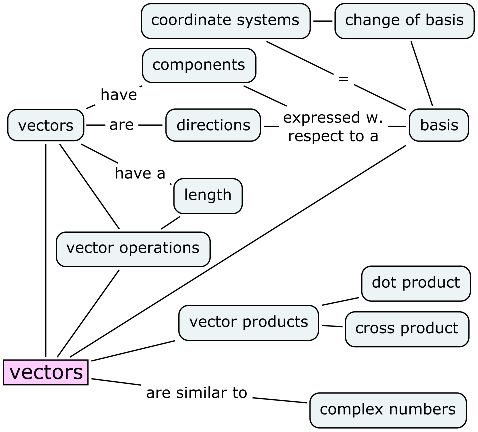


Figure 3.1: This figure illustrates the new concepts related to vectors. As you can see, there is quite a bit of new vocabulary to learn, but don't be phased—all these terms are just fancy ways of talking about arrows.

3.1 Great outdoors

Vectors are directions for getting from point A to point B. Directions can be given in terms of street names and visual landmarks, or with respect to a coordinate system.

While on vacation in British Columbia, you want to visit a certain outdoor location your friend told you about. Your friend isn't available to take you there himself, but he has sent you *directions* for how to get to the place from the bus stop:

```
Sup G. Go to bus stop number 345. Bring a compass.
Walk 2 km north then 3 km east. You will find X there.
```

This text message contains all the information you need to find X.

Act 1: Following directions

You arrive at the bus station, located at the top of a hill. From this height you can see the whole valley, and along the hillside below spreads a beautiful field of tall crops. The crops are so tall they prevent anyone standing in them from seeing too far; good thing you have a compass. You align the compass needle so the red arrow points north. You walk 2 km north, then turn right (east) and walk another 3 km. You arrive at X.

Okay, back to vectors. In this case, the *directions* can be also written as a vector $\vec{d}$, expressed as:

$$\vec{d} = 2\text{km}\, \hat{N} + 3\text{km}\, \hat{E}.$$

This is the mathematical expression that corresponds to the directions "walk 2 km north then 3 km east." Here, $\hat{N}$ is a *direction* and the number in front of the direction tells you the distance to walk in that direction.

Act 2: Equivalent directions

Later during your vacation, you decide to return to the location X. You arrive at the bus stop to find there is a slight problem. From your position, you can see a kilometre to the north, where a group of armed and threatening-looking men stand, waiting to ambush anyone who tries to cross what has now become a trail through the crops. Clearly the word has spread about X and constant visitors have drawn too much attention to the location.

Well, technically speaking, there is no problem at X. The problem lies on the route that starts north and travels through the ambush squad. Can you find an alternate route that leads to X?

```
"Use math, Luke! Use math!"
```

Recall the commutative property of addition for numbers: $a + b = b + a$. Maybe an analogous property holds for vectors? Indeed, this is the case:

$$\vec{d} = 2\text{km}\, \hat{N} + 3\text{km}\, \hat{E} = 3\text{km}\, \hat{E} + 2\text{km}\, \hat{N}.$$

The $\hat{N}$ directions and the $\hat{E}$ directions obey the commutative property. Since the directions can be followed in any order, you can first walk the 3 km east, then walk 2 km north and arrive at X again.

Act 3: Efficiency

It takes 5 km of walking to travel from the bus stop to X, and another 5 km to travel back to the bus stop. Thus, it takes a total of 10 km walking every time you want to go to X. Can you find a quicker route? What is the fastest way from the bus stop to the destination?

Instead of walking in the east and north directions, it would be quicker if you take the diagonal to the destination. Using Pythagoras' theorem you can calculate the length of the diagonal. When the side lengths are 3 and 2, the diagonal has length $\sqrt{3^2+2^2} = \sqrt{9+4} = \sqrt{13} = 3.60555\ldots$. The length of the diagonal route is 3.6 km, which means the diagonal route saves you a whole 1.4 km of walking in each direction.

But perhaps seeking efficiency is not always necessary! You could take a longer path on the way back and give yourself time to enjoy the great outdoors.

Discussion

Vectors are directions for getting from one point to another point. To indicate directions on maps, we use the four cardinal directions: $\hat{N}$, $\hat{S}$, $\hat{E}$, $\hat{W}$. In math, however, we will use only two of the cardinals—$\hat{E} \equiv \hat{x}$ and $\hat{N} \equiv \hat{y}$—since they fit nicely with the usual way of drawing the Cartesian plane. We don't need an $\hat{S}$ direction because we can represent downward distances as negative distances in the $\hat{N}$ direction. Similarly, $\hat{W}$ is the same as negative $\hat{E}$.

From now on, when we talk about vectors we will always represent them with respect to the standard coordinate system $\hat{x}$ and $\hat{y}$, and use *bracket notation*,

$$(v_x, v_y) \equiv v_x\,\hat{x} \;+\; v_y\,\hat{y}.$$

Bracket notation is nice because it's compact, which is good since we will be doing a lot of calculations with vectors. Instead of explicitly writing out all the directions, we will automatically assume that the first number in the bracket is the $\hat{x}$ distance and the second number is the $\hat{y}$ distance.

3.2 Vectors

Vectors are extremely useful in all areas of life. In physics, for example, we use a vector to describe the velocity of an object. It is not sufficient to say that the speed of a tennis ball is 20[m/s]: we must also specify the direction in which the ball is moving. Both of the two velocities

$$\vec{v}_1 = (20, 0) \qquad \text{and} \qquad \vec{v}_2 = (0, 20)$$

describe motion at the speed of 20[m/s]; but since one velocity points along the x-axis, and the other points along the y-axis, they are *completely* different velocities. The velocity vector contains information about the object's speed *and* direction. The direction makes a big difference. If it turns out that the tennis ball is coming your way, you need to get out of the way!

This section's main idea is that **vectors are not the same as numbers**. A vector is a special kind of mathematical object that is *made up of* numbers. Before we begin any calculations with vectors, we need to think about the basic mathematical operations that we can perform on vectors. We will define vector addition $\vec{u} + \vec{v}$, vector subtraction $\vec{u} - \vec{v}$, vector scaling $\alpha\vec{v}$, and other operations. We will also discuss two different notions of *vector product*, which have useful geometrical properties.

Definitions

The two dimensional vector $\vec{v} \in \mathbb{R}^2$ is equivalent to a *pair of numbers* $\vec{v} \equiv (v_x, v_y)$. We call v_x the x-component of $\vec{v}$, and v_y is the y-component of $\vec{v}$.

Vector representations

We'll use three equivalent ways to denote vectors:

- $\vec{v} = (v_x, v_y)$: component notation, where the vector is represented as a pair of coordinates with respect to the x-axis and the y-axis
- $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath}$: unit vector notation, where the vector is expressed in terms of the unit vectors $\hat{\imath} = (1, 0)$ and $\hat{\jmath} = (0, 1)$
- $\vec{v} = \|\vec{v}\|\angle\theta$: length-and-direction notation, where the vector is expressed in terms of its *length* $\|\vec{v}\|$ and the angle θ that the vector makes with the x-axis.

These three notations describe different aspects of vectors, and we will use them throughout the rest of the book. We'll learn how to convert between them—both algebraically (with pen, paper, and calculator) and intuitively (by drawing arrows).

Vector operations

Consider two vectors, $\vec{u} = (u_x, u_y)$ and $\vec{v} = (v_x, v_y)$, and assume that $\alpha \in \mathbb{R}$ is an arbitrary constant. The following operations are defined for these vectors:

- **Addition:** $\vec{u} + \vec{v} = (u_x + v_x, u_y + v_y)$
- **Subtraction:** $\vec{u} - \vec{v} = (u_x - v_x, u_y - v_y)$
- **Scaling:** $\alpha\vec{u} = (\alpha u_x, \alpha u_y)$
- **Dot product:** $\vec{u} \cdot \vec{v} = u_x v_x + u_y v_y$
- **Length:** $\|\vec{u}\| = \sqrt{\vec{u} \cdot \vec{u}} = \sqrt{u_x^2 + u_y^2}$. We will also sometimes simply use the letter u to denote the length of $\vec{u}$.
- **Cross product:** $\vec{u} \times \vec{v} = (u_y v_z - u_z v_y,\ u_z v_x - u_x v_z,\ u_x v_y - u_y v_x)$. The cross product is only defined for three-dimensional vectors like $\vec{u} = (u_x, u_y, u_z)$ and $\vec{v} = (v_x, v_y, v_z)$.

Pay careful attention to the dot product and the cross product. Although they're called products, these operations behave much differently than taking the product of two numbers. Also note, there is no notion of vector division.

Vector algebra

Addition and subtraction Just like numbers, you can add vectors

$$\vec{v} + \vec{w} = (v_x, v_y) + (w_x, w_y) = (v_x + w_x, v_y + w_y),$$

subtract them

$$\vec{v} - \vec{w} = (v_x, v_y) - (w_x, w_y) = (v_x - w_x, v_y - w_y),$$

and solve all kinds of equations where the unknown variable is a vector. This is not a formidably complicated new development in mathematics. Performing arithmetic calculations on vectors simply requires **carrying out arithmetic operations on their components**. Given two vectors, $\vec{v} = (4, 2)$ and $\vec{w} = (3, 7)$, their difference is computed as $\vec{v} - \vec{w} = (4, 2) - (3, 7) = (1, -5)$.

Scaling We can also *scale* a vector by any number $\alpha \in \mathbb{R}$:

$$\alpha\vec{v} = (\alpha v_x, \alpha v_y),$$

where each component is multiplied by the scaling factor α. Scaling changes the length of a vector. If $\alpha > 1$ the vector will get longer, and if $0 \leq \alpha < 1$ then the vector will become shorter. If α is a negative number, the scaled vector will point in the opposite direction.

Length A vector's length is obtained from Pythagoras' theorem. Imagine a triangle with one side of length v_x and the other side of length v_y; the length of the vector is equal to the length of the triangle's hypotenuse:

$$\|\vec{v}\|^2 = v_x^2 + v_y^2 \qquad \Rightarrow \qquad \|\vec{v}\| = \sqrt{v_x^2 + v_y^2}\,.$$

A common technique is to scale a vector $\vec{v}$ by the inverse of its length $\frac{1}{\|\vec{v}\|}$ to obtain a unit-length vector that points in the same direction as $\vec{v}$:

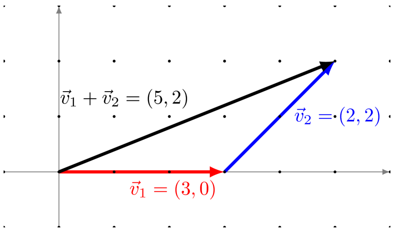
$$\hat{v} \equiv \frac{\vec{v}}{\|\vec{v}\|} = \left(\frac{v_x}{\|\vec{v}\|}, \frac{v_y}{\|\vec{v}\|} \right).$$

Unit vectors (denoted with a hat instead of an arrow) are useful when you want to describe only a direction in space without any specific length in mind. Verify that $\|\hat{v}\| = 1$.

Vector as arrows

So far, we described how to perform algebraic operations on vectors in terms of their components. Vector operations can also be interpreted geometrically, as operations on two-dimensional arrows in the Cartesian plane.

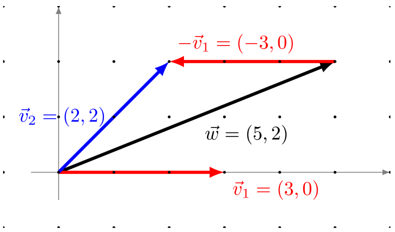
Vector addition The sum of two vectors corresponds the combined displacement of the two vectors. The diagram on the right illustrates the addition of two vectors, $\vec{v}_1 = (3,0)$ and $\vec{v}_2 = (2,2)$. The sum of the two vectors is the vector $\vec{v}_1 + \vec{v}_2 = (3,0) + (2,2) = (5,2)$.



Vector subtraction Before we describe vector subtraction, note that multiplying a vector by a scaling factor $\alpha = -1$ gives a vector of the same length as the original, but pointing in the opposite direction.

This fact is useful if you want to subtract two vectors using the graphical approach. Subtracting a vector is the same as adding the negative of the vector:

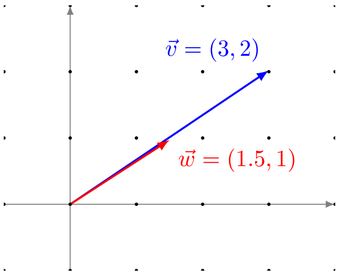
$$\vec{w} - \vec{v}_1 = \vec{w} + (-\vec{v}_1) = \vec{v}_2.$$



The diagram on the right illustrates the graphical procedure for subtracting the vector $\vec{v}_1 = (3,0)$ from the vector $\vec{w} = (5,2)$. Subtraction of $\vec{v}_1 = (3,0)$ is the same as addition of $-\vec{v}_1 = (-3,0)$.

Scaling The scaling operation acts to change the length of a vector. Suppose we want to obtain a vector in the same direction as the vector $\vec{v} = (3, 2)$, but half as long. "Half as long" corresponds to a scaling factor of $\alpha = 0.5$. The scaled-down vector is $\vec{w} = 0.5\vec{v} = (1.5, 1)$.

Conversely, we can think of the vector $\vec{v}$ as being twice as long as the vector $\vec{w}$.

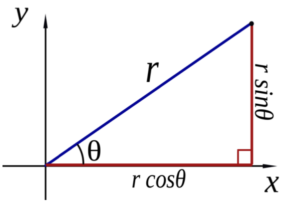


Length and direction representation

So far, we've seen how to represent a vector in terms of its components. There is also another way of representing vectors: we can specify a vector in terms of its length $||\vec{v}||$ and its direction—the angle it make with the x-axis. For example, the vector $(1, 1)$ can also be written as $\sqrt{2}\angle 45°$. This magnitude-and-direction notation is useful because it makes it easy to see the "size" of vectors. On the other hand, vector arithmetic operations are much easier to carry out in the component notation. We will use the following formulas for converting between the two notations.

To convert the length-and-direction vector $\|\vec{r}\|\angle\theta$ into an x-component and a y-component (r_x, r_y), use the formulas

$$r_x = \|\vec{r}\| \cos\theta \quad \text{and} \quad r_y = \|\vec{r}\| \sin\theta.$$



To convert from component notation (r_x, r_y) to length-and-direction $\|\vec{r}\|\angle\theta$, use

$$r = \|\vec{r}\| = \sqrt{r_x^2 + r_y^2} \qquad \text{and} \qquad \theta = \tan^{-1}\left(\frac{r_y}{r_x}\right).$$

Note that the second part of the equation involves the inverse tangent function. By convention, the function $\tan^{-1}$ returns values between $\pi/2$ $(90°)$ and $-\pi/2$

($-90°$). You must be careful when finding the θ of vectors with an angle outside of this range. Specifically, for vectors with $v_x < 0$, you must add π ($180°$) to $\tan^{-1}(r_y/r_x)$ to obtain the correct θ.

Unit vector notation

As discussed above, we can think of a vector $\vec{v} = (v_x, v_y, v_z)$ as a command to "go a distance v_x in the x-direction, a distance v_y in the y-direction, and v_z in the z-direction."

To write this set of commands more explicitly, we can use multiples of the vectors $\hat{\imath}, \hat{\jmath}$, and $\hat{k}$. These are the unit vectors pointing in the x, y, and z directions, respectively:

$$\hat{\imath} = (1,0,0), \qquad \hat{\jmath} = (0,1,0), \quad \text{and} \quad \hat{k} = (0,0,1).$$

Any number multiplied by $\hat{\imath}$ corresponds to a vector with that number in the first coordinate. For example, $3\hat{\imath} \equiv (3,0,0)$. Similarly, $4\hat{\jmath} \equiv (0,4,0)$ and $5\hat{k} \equiv (0,0,5)$.

In physics, we tend to perform a lot of numerical calculations with vectors; to make things easier, we often use unit vector notation:

$$v_x\hat{\imath} + v_y\hat{\jmath} + v_z\hat{k} \qquad \Leftrightarrow \qquad \vec{v} \qquad \Leftrightarrow \qquad (v_x, v_y, v_z).$$

The addition rule remains the same for the new notation:

$$\underbrace{2\hat{\imath} + 3\hat{\jmath}}_{\vec{v}} \; + \; \underbrace{5\hat{\imath} - 2\hat{\jmath}}_{\vec{w}} \; = \; \underbrace{7\hat{\imath} + 1\hat{\jmath}}_{\vec{v}+\vec{w}}.$$

It's the same story repeating all over again: we need to add $\hat{\imath}$s with $\hat{\imath}$s, and $\hat{\jmath}$s with $\hat{\jmath}$s.

Examples

Simple example

Compute the sum $\vec{s} = 4\hat{\imath} + 5\angle 30°$. Express your answer in the length-and-direction notation.

Since we want to carry out an addition, and since addition is performed in terms of the components, our first step is to convert $5\angle 30^\circ$ into component notation. We find $5\angle 30^\circ = (5\cos 30^\circ)\hat{\imath} + (5\sin 30^\circ)\hat{\jmath} = 5\frac{\sqrt{3}}{2}\hat{\imath} + \frac{5}{2}\hat{\jmath}$. We can now compute the sum:

$$\vec{s} \;=\; 4\hat{\imath} \;+\; 5\tfrac{\sqrt{3}}{2}\hat{\imath} + \tfrac{5}{2}\hat{\jmath} \;=\; (4 + 5\tfrac{\sqrt{3}}{2})\hat{\imath} + (\tfrac{5}{2})\hat{\jmath}\,.$$

The x-component of the sum is $s_x = (4 + 5\frac{\sqrt{3}}{2})$, and the y-component of the sum is $s_y = (\frac{5}{2})$. To express the answer as a length and a direction, we compute the length $\|\vec{s}\| = \sqrt{s_x^2 + s_y^2} = 8.697$ and the direction $\tan^{-1}(s_y/s_x) = 16.7^\circ$. The answer is $\vec{s} = 8.697\angle 16.7^\circ$.

Vector addition example

You're heading to physics class after a "safety meeting" with a friend, and are looking forward to two hours of finding absolute amazement and awe in the laws of Mother Nature. As it turns out, there is no enlightenment to be had that day because there is going to be an in-class midterm. The first question involves a block sliding down an incline. You look at it, draw a little diagram, and then wonder how the hell you are going to find the net force acting on the block. The three forces acting on the block are $\vec{W} = 30\angle -90^\circ$, $\vec{N} = 200\angle -290^\circ$, and $\vec{F}_f = 50\angle 60^\circ$.

You happen to remember the net force formula:

$$\sum \vec{F} = \vec{F}_{\text{net}} = m\vec{a} \qquad \text{[Newton's } 2^{\text{nd}} \text{ law]}.$$

You get the feeling Newton's 2nd law is the answer to all your troubles. You sense this formula is certainly the key because you saw the keyword "net force" when reading the question, and notice "net force" also appears in this very equation.

The net force is the sum of all forces acting on the block:

$$\vec{F}_{\text{net}} = \sum \vec{F} = \vec{W} + \vec{N} + \vec{F}_f.$$

All that separates you from the answer is the addition of these vectors. Vectors have components, and there is the whole sin/cos procedure for decomposing length-

and-direction vectors in terms of their components. If you have the vectors as components you will be able to add them and find the net force.

Okay, chill! Let's do this one step at a time. The net force must have an x-component, which, according to the equation, must equal the sum of the x-components of all the forces:

$$\begin{aligned}F_{\text{net},x} &= W_x + N_x + F_{f,x}\\ &= 30\cos(-90^\circ) + 200\cos(-290^\circ) + 50\cos(60^\circ)\\ &= 93.4.\end{aligned}$$

Now find the y-component of the net force using the $\sin$ of the angles:

$$\begin{aligned}F_{\text{net},y} &= W_y + N_y + F_{f,y}\\ &= 30\sin(-90^\circ) + 200\sin(-290^\circ) + 50\sin(60^\circ)\\ &= 201.2.\end{aligned}$$

Combining the two components of the vector, we get the final answer:

$$\vec{F}_{\text{net}} = (F_{\text{net},x}, F_{\text{net},y}) = (93.4, 201.2) = 93.4\hat{\imath} + 201.2\hat{\jmath}.$$

Bam! Just like that you're done, because you overstand them vectors!

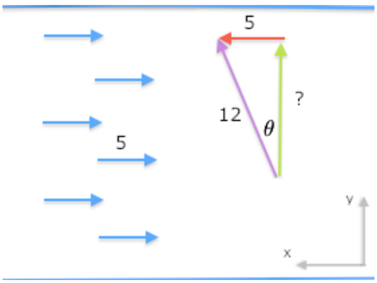
Relative motion example

A boat can reach a top speed of 12 knots in calm seas. Instead of cruising through a calm sea, however, the boat's crew is trying to sail up the St-Laurence river. The speed of the current is 5 knots.

If the boat travels directly upstream at full throttle $12\vec{\imath}$, then the speed of the boat relative to the shore will be

$$12\hat{\imath} - 5\hat{\imath} = 7\hat{\imath},$$

since we have to "deduct" the speed of the current from the speed of the boat relative to the water.



If the crew wants to cross the river perpendicular to the current flow, they can use some of the boat's thrust to counterbalance the current, and the remaining thrust to push across. In what direction should the boat sail to cross the river? We are looking for the direction of $\vec{v}$ the boat should take such that, after adding in the velocity of the current, the boat moves in a straight line between the two banks (the $\hat{\jmath}$ direction).

A picture is necessary: draw a river, then draw a triangle in the river with its long leg perpendicular to the current flow. Make the short leg of length 5. We will take the up-the-river component of the speed $\vec{v}$ to be equal to $5\hat{\imath}$, so that it cancels exactly the $-5\hat{\imath}$ flow of the river. Finally, label the hypotenuse with length 12, since this is the speed of the boat relative to the surface of the water.

From all of this we can answer the question like professionals. You want the angle? Well, we have that $12\sin(\theta) = 5$, where θ is the angle of the boat's course relative to the straight line between the two banks. We can use the inverse-sin function to solve for the angle:

$$\theta = \sin^{-1}\left(\frac{5}{12}\right) = 24.62^\circ.$$

The across-the-river component speed can be calculated from $v_y = 12\cos(\theta) = 10.91$, or from Pythagoras' theorem if you prefer $v_y = \sqrt{\|\vec{v}\|^2 - v_x^2} = \sqrt{12^2 - 5^2} = 10.91$.

Discussion

Vector dimensions

The most common types of vectors are two-dimensional vectors (like the ones in the Cartesian plane), and three-dimensional vectors (directions in 3D space). 2D and 3D vectors are easier to work with because we can visualize them and draw them in diagrams. In general, vectors can exist in any number of dimensions. An example of a n-dimensional vector is

$$\vec{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n.$$

The rules of vector algebra apply in higher dimensions, but our ability to visualize stops at three dimensions.

Coordinate system

The geometrical interpretation of vectors depends on the coordinate system in which the vectors are represented. Throughout this section we have used the x, y, and z axes, and we've described vectors as components along each of these directions. This is a very convenient coordinate system; we have a set of three *perpendicular* axes, and a set of three unit vectors $\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$ that point along each of the three axis directions. Every vector is implicitly defined in terms of this coordinate system. When you and I talk about the vector $\vec{v} = 3\hat{\imath} + 4\hat{\jmath} + 2\hat{k}$, we are really saying, "start from the origin $(0,0,0)$, move 3 units in the x-direction, then move 4 units in the y-direction, and finally move 2 units in the z-direction." It is simpler to express these directions as $\vec{v} = (3,4,2)$, while remembering that the numbers in the bracket measure distances *relative* to the xyz-coordinate system.

It turns out, using the xyz-coordinate system and the vectors $\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$ is just one of many possible ways we can represent vectors. We can represent a vector $\vec{v}$ as coefficients (v_1, v_2, v_3) with respect to any *basis* $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ as follows: $\vec{v} = v_1\hat{e}_1 + v_2\hat{e}_2 + v_3\hat{e}_3$. What is a basis, you ask? I'm glad you asked, because this is the subject of the next section.

3.3 Basis

One of the most important concepts in the study of vectors is the concept of a basis. Consider the space of three-dimensional vectors $\mathbb{R}^3$. A *basis* for $\mathbb{R}^3$ is a set of vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ which can be used as a coordinate system for $\mathbb{R}^3$. If the set of vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ is a basis, then you can *represent* any vector $\vec{v} \in \mathbb{R}^3$ as coefficients (v_1, v_2, v_3) *with respect to* that basis:

$$\vec{v} = v_1\hat{e}_1 + v_2\hat{e}_2 + v_3\hat{e}_3.$$

The vector $\vec{v}$ is obtained by measuring out a distance v_1 in the $\hat{e}_1$ direction, a distance v_2 in the $\hat{e}_2$ direction, and a distance v_3 in the $\hat{e}_3$ direction.

You are already familiar with the *standard* basis $\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$, which is associated with the xyz-coordinate system. You know that any vector $\vec{v} \in \mathbb{R}^3$ can be expressed as a triplet (v_x, v_y, v_z) with respect to the basis $\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$ through the formula $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath} + v_z\hat{k}$. In this section, we'll discuss how to represent vectors with respect to other bases.

An analogy

Let's start with a simple example of a basis. If you look at the HTML code behind any web page, you're sure to find at least one mention of the colour stylesheet directive such as `background-color:#336699;`. The numbers should be interpreted as a triplet of values $(33, 66, 99)$, each value describing the amount of red, green, and blue needed to create a given colour. Let us call the colour described by the triplet $(33, 66, 99)$ CoolBlue. This convention for colour representation is called the RGB colour model and we can think of it as the *RGB basis*. A basis is a set of elements that can be combined together to express something more complicated. In our case, the **R**, **G**, and **B** elements are pure colours that can create any colour when mixed appropriately. Schematically, we can write this mixing idea as

$$\text{CoolBlue} = (33, 66, 99)_{RGB} = 33\mathbf{R} + 66\mathbf{G} + 99\mathbf{B},$$

where the *coefficients* determine the strength of each colour component. To create the colour, we combine its components as symbolized by the $+$ operation.

The cyan, magenta, and yellow (CMY) colour model is another basis for representing colours. To express the "cool blue" colour in the CMY basis, you will need the following coefficients:

$$(33,66,99)_{RGB} = \text{CoolBlue} = (222,189,156)_{CMY} = 222\mathbf{C} + 189\mathbf{M} + 156\mathbf{Y}.$$

The *same* colour CoolBlue is represented by a *different* set of coefficients when the CMY colour basis is used.

Note that a triplet of coefficients by itself does not mean anything unless we know the basis being used. For example, if we were to interpret the triplet of coordinates $(33,66,99)$ with respect to the CMY basis, will would obtain a completely different colour, which would not be cool at all.

A basis is required to convert mathematical objects like the triplet (a,b,c) into real-world ideas like colours. As exemplified above, to avoid any ambiguity we can use a subscript after the bracket to indicate the basis associated with each triplet of coefficients.

Discussion

It's hard to over-emphasize the importance of the basis—the coordinate system you will use to describe vectors. The choice of coordinate system is the bridge between real-world vector quantities and their mathematical representation in terms of components. Every time you solve a problem with vectors, **the first thing you should do is draw a coordinate system**, and think of vector components as measuring out a distance along this coordinate system.

3.4 Vector products

If addition of two vectors $\vec{v}$ and $\vec{w}$ corresponds to the addition of their components $(v_x+w_x, v_y+w_y, v_z+w_z)$, you might logically think that the product of two vectors will correspond to the product of their components (v_xw_x, v_yw_y, v_zw_z), however, this way of multiplying vectors is not used in practice. Instead, we use the dot product and the cross product.

The *dot product* tells you how similar two vectors are to each other:

$$\vec{v} \cdot \vec{w} \equiv v_xw_x + v_yw_y + v_zw_z \equiv \|\vec{v}\|\|\vec{w}\|\cos(\varphi) \quad \in \mathbb{R},$$

where φ is the angle between the two vectors. The factor $\cos(\varphi)$ is largest when the two vectors point in the same direction because the angle between them will be $\varphi = 0$ and $\cos(0) = 1$.

The exact formula for the *cross product* is more complicated so I will not show it to you just yet. What is important to know is that the cross product of two vectors is another vector:

$$\vec{v} \times \vec{w} = \{ \text{ a vector perpendicular to both } \vec{v} \text{ and } \vec{w} \ \} \quad \in \mathbb{R}^3.$$

If you take the cross product of one vector pointing in the x-direction with another vector pointing in the y-direction, the result will be a vector in the z-direction.

Dot product

The *dot product* between two vectors can be computed using either the algebraic formula

$$\vec{v} \cdot \vec{w} \equiv v_x w_x + v_y w_y + v_z w_z,$$

or the geometrical formula

$$\vec{v} \cdot \vec{w} \equiv \|\vec{v}\| \|\vec{w}\| \cos(\varphi),$$

where φ is the angle between the two vectors. This operation is also known as the *inner* product or *scalar* product. The name *scalar* comes from the fact that the result of the dot product is a scalar number—a number that does not change when the basis changes.

The signature for the dot product operation is

$$\cdot : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}.$$

The dot product takes two vectors as inputs and produces a real number as output.

The geometric factor $\cos(\varphi)$ depends on the relative orientation of the two vectors as follows:

- If the vectors point in the same direction, then $\cos(\varphi) = \cos(0°) = 1$ and so $\vec{v} \cdot \vec{w} = \|\vec{v}\|\|\vec{w}\|$.
- If the vectors are perpendicular to each other, then $\cos(\varphi) = \cos(90°) = 0$ and so $\vec{v} \cdot \vec{w} = \|\vec{v}\|\|\vec{w}\|(0) = 0$.
- If the vectors point in exactly opposite directions, then $\cos(\varphi) = \cos(180°) = -1$ and so $\vec{v} \cdot \vec{w} = -\|\vec{v}\|\|\vec{w}\|$.

Cross product

The *cross product* takes two vectors as inputs and produces another vector as the output:

$$\times : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3.$$

Because the output of this operation is a vector, we sometimes refer to the cross product as the *vector* product.

The cross products of individual basis elements are defined as follows:

$$\hat{\imath} \times \hat{\jmath} = \hat{k}, \quad \hat{\jmath} \times \hat{k} = \hat{\imath}, \quad \hat{k} \times \hat{\imath} = \hat{\jmath}.$$

The cross product is *anti-symmetric* in its inputs, which means swapping the order of the inputs introduces a negative sign in the output:

$$\hat{\jmath} \times \hat{\imath} = -\hat{k}, \quad \hat{k} \times \hat{\jmath} = -\hat{\imath}, \quad \hat{\imath} \times \hat{k} = -\hat{\jmath}.$$

I bet you had never seen a product like this before. Most likely, the products you've seen in math have been *commutative*, which means the order of the inputs doesn't matter. The product of two numbers is commutative $ab = ba$, and the dot product is commutative $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$, but the cross product of two vectors is *non-commutative* $\hat{\imath} \times \hat{\jmath} \neq \hat{\jmath} \times \hat{\imath}$.

For two arbitrary vectors $\vec{a} = (a_x, a_y, a_z)$ and $\vec{b} = (b_x, b_y, b_z)$, the cross product is calculated as

$$\vec{a} \times \vec{b} = (a_y b_z - a_z b_y,\ a_z b_x - a_x b_z,\ a_x b_y - a_y b_x)\,.$$

The cross product's output has a length that is proportional to the $\sin$ of the angle between the vectors:

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

The direction of the vector $(\vec{a} \times \vec{b})$ is perpendicular to both $\vec{a}$ and $\vec{b}$.

3.5 Complex numbers

By now, you've heard about complex numbers $\mathbb{C}$. The word "complex" is an intimidating word. Surely it must be a complex task to learn about the complex numbers. That may be true in general, but it helps if you know about vectors. Complex numbers are similar to two-dimensional vectors $\vec{v} \in \mathbb{R}^2$. We add and subtract complex numbers like vectors. Complex numbers also have components, length, and "direction." If you understand vectors, you will understand complex numbers at almost no additional mental cost.

We'll begin with a practical problem.

Example

Suppose you are asked to solve the following quadratic equation:

$$x^2 + 1 = 0.$$

You're looking for a number x, such that $x^2 = -1$. If you are only allowed to give real answers (the set of real numbers is denoted $\mathbb{R}$), then there is no answer to this question. In other words, this equation has no solutions. Graphically speaking, this is because the quadratic function $f(x) = x^2 + 1$ does not cross the x-axis.

However, we're not going to take nothing as an answer. We will imagine a new number called i that satisfies $i^2 = -1$. We call i the unit imaginary number. The solutions to the equation are therefore $x_1 = i$ and $x_2 = -i$. There are two solutions because the equation was quadratic. We can check that $i^2 + 1 = -1 + 1 = 0$ and also $(-i)^2 + 1 = (-1)^2 i^2 + 1 = i^2 + 1 = 0$.

Thus, while the equation $x^2+1 = 0$ has no real solutions, it *does* have solutions if we allow the answers to be complex numbers.

Definitions

Complex numbers have a real part and an imaginary part:

- i: the unit imaginary number $i \equiv \sqrt{-1}$ or $i^2 = -1$
- bi: an imaginary number that is equal to b times i
- $\mathbb{R}$: the set of real numbers
- $\mathbb{C}$: the set of complex numbers $\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}$
- $z = a + bi$: a complex number
- $\text{Re}\{z\} = a$: the real part of z
- $\text{Im}\{z\} = b$: the imaginary part of z
- $\bar{z}$: the *complex conjugate* of z. If $z = a + bi$, then $\bar{z} = a - bi$.

The polar representation of complex numbers:

- $z = |z|\angle\phi_z = |z|\cos\phi_z + i|z|\sin\phi_z$
- $|z| = \sqrt{\bar{z}z} = \sqrt{a^2 + b^2}$: the *magnitude* of $z = a + bi$
- $\phi_z = \tan^{-1}(b/a)$: the phase of $z = a + bi$
- $\text{Re}\{z\} = |z|\cos\phi_z$
- $\text{Im}\{z\} = |z|\sin\phi_z$

Formulas

Addition and subtraction

Just as we performed the addition of vectors component by component, we perform addition on complex numbers by adding the real parts together and adding the imaginary part together:

$$(a + bi) + (c + di) = (a + c) + (b + d)i.$$

Polar representation

We can give a geometrical interpretation of the complex numbers by extending the real number line into a two-dimensional plane called the *complex plane*. The horizontal axis in the complex plane measures the *real* part of the number. The vertical axis measures the *imaginary* part. Complex numbers are vectors in the complex plane.

It is possible to represent any complex number $z = a + bi$ in terms of its *magnitude* and its *phase*:

$$z = |z|\angle\phi_z = (|z|\cos\phi_z) + (|z|\sin\phi_z)i.$$

The magnitude of a complex number $z = a + bi$ is

$$|z| = \sqrt{a^2 + b^2}.$$

This corresponds to the *length* of the vector which represents the complex number in the complex plane. The formula is obtained by using Pythagoras' theorem.

The *phase* of the complex number is:

$$\phi_z = \tan^{-1}(b/a).$$

The phase corresponds to the angle z forms with the real axis.

Multiplication

The product of two complex numbers is computed using the usual rules of algebra:

$$(a+bi)(c+di) = (ac-bd)+(ad+bc)i.$$

In the polar representation, the product is

$$(p\angle\phi)(q\angle\psi) = pq\angle(\phi+\psi).$$

Cardano's example

One of the earliest examples of reasoning involving complex numbers was given by Gerolamo Cardano in his 1545 book *Ars Magna*. Cardano wrote, "If someone says to you, divide 10 into two parts, one of which multiplied into the other shall produce 40, it is evident that this case or question is impossible." We want to find numbers x_1 and x_2 such that $x_1+x_2=10$ and $x_1x_2=40$. This sounds kind of impossible. Or is it?

"Nevertheless," Cardano said, "we shall solve it in this fashion:

$$x_1 = 5+\sqrt{15}i \text{ and } x_2 = 5-\sqrt{15}i.\text{"}$$

When you add x_1+x_2 you obtain 10. When you multiply the two numbers the answer is

$$x_1x_2 = \left(5+\sqrt{15}i\right)\left(5-\sqrt{15}i\right) = 25-\sqrt{15}^2i^2 = 25+15 = 40.$$

Hence $5+\sqrt{15}i$ and $5-\sqrt{15}i$ are two numbers whose sum is 10 and whose product is 40.

Example

Both i and -1 have a magnitude of 1. i has phase $\frac{\pi}{2}$ (90°), while -1 has phase π (180°). Consider the product of these two numbers:

$$(i)(-1) \;=\; (1\angle\tfrac{\pi}{2})(1\angle\pi) \;=\; 1\angle\tfrac{3\pi}{2} \;=\; -i.$$

Multiplication by i is effectively a rotation by 90 degrees leftward.

Division

Let's look at the procedure for dividing complex numbers:

$$\frac{(a+bi)}{(c+di)} = \frac{(a+bi)}{(c+di)}\frac{(c-di)}{(c-di)} = (a+bi)\frac{(c-di)}{(c^2+d^2)} = (a+bi)\frac{\overline{c+di}}{|c+di|^2}.$$

In other words, to divide the number z by the complex number s, compute $\bar{s}$ and $|s|^2 = s\bar{s}$ and then use

$$z/s = z\frac{\bar{s}}{|s|^2}.$$

You can think of $\frac{\bar{s}}{|s|^2}$ as being equivalent to s^{-1}.

Fundamental theorem of algebra

The solutions to *any* polynomial equation $a_0 + a_1x + \cdots a_nx^n = 0$ are of the form

$$z = a + bi.$$

In other words, any polynomial $P(x)$ of n^{th} degree can be written as

$$P(x) = (x - z_1)(x - z_2)\cdots(x - z_n),$$

where $z_i \in \mathbb{C}$ are the polynomial's *complex* roots.

Before today, you might have said the equation $x^2 + 1 = 0$ has no solutions. Now you know its solutions are the complex numbers $z_1 = i$ and $z_2 = -i$.

Euler's formula

You already know $\cos\theta$ is a shifted version of $\sin\theta$, so it's clear these two functions are related. It turns out the exponential function is also related to $\sin$ and $\cos$. Lo and behold, we have Euler's formula:

$$e^{i\theta} = \cos\theta + i\sin\theta\,.$$

Inputting an imaginary number to the exponential function outputs a complex number that contains both $\cos$ and $\sin$. Euler's formula gives us an alternate notation for the polar representation of complex numbers: $z = |z|\angle\phi_z = |z|e^{i\phi_z}$.

If you want to impress your friends with your math knowledge, plug $\theta = \pi$ into the above equation to find

$$e^{i\pi} = \cos(\pi) + i\sin(\pi) = -1,$$

which can be rearranged into the form, $e^{\pi i} + 1 = 0$. This equation shows a relationship between the five most important numbers in all of mathematics: Euler's number $e = 2.71828\ldots$, $\pi = 3.14159\ldots$, the imaginary number i, 1, and zero. It's kind of cool to see all these important numbers reunited in one equation, don't you agree?

De Moivre's theorem

By replacing θ in Euler's formula with $n\theta$, we obtain de Moivre's theorem:

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

De Moivre's Theorem makes sense if you think of the complex number $z = e^{i\theta} = \cos\theta + i\sin\theta$, raised to the n^{th} power:

$$(\cos\theta + i\sin\theta)^n = z^n = (e^{i\theta})^n = e^{in\theta} = \cos n\theta + i\sin n\theta.$$

Setting $n = 2$ in de Moivre's formula, we can derive the double angle formulas as the real and imaginary parts of the following equation:

$$(\cos^2\theta - \sin^2\theta) + (2\sin\theta\cos\theta)i = \cos(2\theta) + \sin(2\theta)i.$$

Links

[Mini tutorial on the complex numbers]
http://paste.lisp.org/display/133628

Chapter 4

Mechanics

4.1 Introduction

Mechanics is the precise study of the motion of objects, the forces acting on them, and more abstract concepts such as momentum and energy. You probably have an intuitive understanding of these concepts already. In this chapter we will learn how to use precise mathematical equations to support your intuition.

Newton's laws

Mechanics is the part of physics that is easiest to understand. Starting from three general principles known as *Newton's laws*, we can figure out pretty much everything about the motion of objects.

Newton's three laws of motion:

1. In the absence of external forces, objects will maintain their velocity and their direction of motion.
2. A force acting on an object causes an acceleration inversely proportional to the mass of the object: $\vec{F} = m\vec{a}$.
3. For each force $\vec{F}_{12}$ applied by Object 1 on Object 2, there is an equal and opposite force $\vec{F}_{21}$ that Object 2 exerts on Object 1.

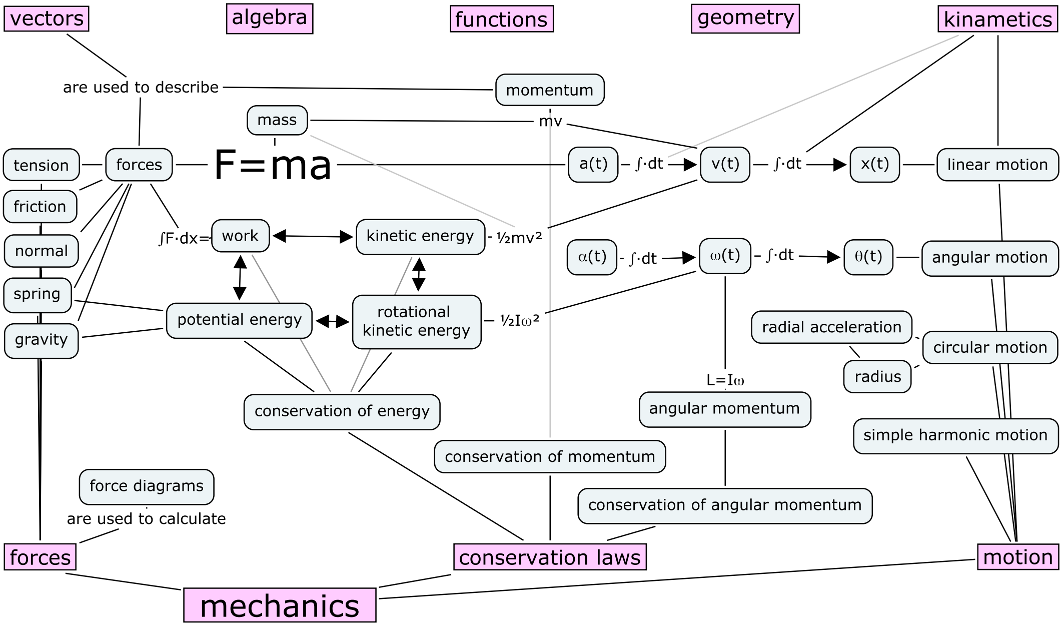


Figure 4.1: The concepts of mechanics. Forces are the cause of motion. We can also analyze the motion of objects in terms of the concepts of energy and momentum. If you understand the connections between all of the above concepts, you understand mechanics.

The cool part of learning physics is that it teaches us to think about the laws of nature in terms of simple principles. Complicated phenomena can be broken down and understood in terms of basic theories. The laws of physics can be expressed in terms of mathematical equations. There are about 20 such equations (see page 465 in the back of the book). In this chapter you will learn how to use these equations to solve all kinds of physics problems.

Kinematics is the study of motion

To solve a physics problem is to obtain the *equation of motion* $x(t)$, which describes the position of the object as a function of time. Once you know $x(t)$, you can answer any question pertaining to the object's motion. To find the initial position x_i of the object, plug $t = 0$ into the equation of motion $x_i = x(0)$. To find the time(s) when the object reaches a certain distance, let's say 20[m] from the origin, solve for t in $x(t) = 20$[m]. Many of the problems on the mechanics final exam will be of this kind, so if you know how to find $x(t)$, you'll be in good shape to ace the exam.

In Chapter 2, we learned about the kinematics of objects moving in one dimension. More specifically, we used integration to obtain the velocity function of an object starting from the knowledge of its acceleration. Integrating the velocity function, we obtain its position function:

$$a(t) \quad \overset{v_i+\int dt}{\longrightarrow} \quad v(t) \quad \overset{x_i+\int dt}{\longrightarrow} \quad x(t).$$

Okay, but how do we obtain the acceleration?

Dynamics is the study of forces

The first step toward finding $x(t)$ is to calculate all the *forces* that act on the object. The forces are the *cause* of the object's acceleration. Newton's second law $F = ma$ states that **a force acting on an object produces an acceleration inversely proportional to the mass of the object**. There are many kinds of forces: the weight of an object $\vec{W}$ is a type of force, the force of friction $\vec{F}_f$ is another type of force, the tension in a rope $\vec{T}$ is yet another type of force, and so on. Note the little arrow on top of each force, which is there to remind you that forces are *vector quantities*. To find the *net force* acting on the object, calculate the sum of all the forces acting on the object $\vec{F}_{\text{net}} \equiv \sum \vec{F}$.

Once you know the net force, you can use the formula $\vec{a}(t) = \frac{\vec{F}_{\text{net}}}{m}$ to find the object's acceleration. Once you know the acceleration $a(t)$, you can compute $x(t)$ using the calculus steps we learned in Chapter 2. The entire procedure for predicting the motion of objects can be summarized as

$$\frac{1}{m}\underbrace{\left(\sum \vec{F} = \vec{F}_{\text{net}}\right)}_{\text{dynamics}} = \underbrace{a(t) \quad \overset{v_i+\int dt}{\longrightarrow} \quad v(t) \quad \overset{x_i+\int dt}{\longrightarrow} \quad x(t)}_{\text{kinematics}}.$$

If you understand the above equation, then you understand mechanics. The goal of this chapter is to introduce you to all the concepts that appear in this equation and explore the relationships between them.

Other stuff

In addition to dynamics and kinematics, this chapter covers a number of other physics topics.

Newton's second law can also be applied to the study of objects in rotation. Angular motion is described by the angle of rotation $\theta(t)$, the angular velocity $\omega(t)$, and the angular acceleration $\alpha(t)$. Angular acceleration is caused by angular force, which we call *torque* $\mathcal{T}$. The principles behind circular motion are almost exactly the same as the principles of linear motion; the only difference being we use *angular* quantities to describe circular motion—instead of describing the motion in terms of [m], [m/s], and [m/s^2], we describe angular motion in terms of [radians], [radians/s], and [radians/s^2].

During a collision between two objects, the sudden spike in the contact force between them can be difficult to measure and quantify. It is therefore not possible to use Newton's law $F = ma$ to find the accelerations of the objects occurring during collisions and predict the motion of the objects using the kinematics approach described above.

To predict the motion of objects after a collision, we can use a *momentum* calculation. An object of mass m moving with velocity $\vec{v}$ has momentum $\vec{p} \equiv m\vec{v}$. The principle of conservation of momentum states that **the total amount of momentum in a system before and after a collision is conserved**. Thus, if two objects with initial momenta $\vec{p}_{i1}$ and $\vec{p}_{i2}$ collide, the total momentum before the collision must be equal to the total momentum after the collision:

$$\sum \vec{p}_i = \sum \vec{p}_f \qquad \Rightarrow \qquad \vec{p}_{i1} + \vec{p}_{i2} = \vec{p}_{f1} + \vec{p}_{f2}.$$

We use this equation to calculate the final momenta $\vec{p}_{f1}$, $\vec{p}_{f2}$ of the objects after the collision.

There is another way to solve physics problems by applying the concept of *energy*. Instead of trying to describe the entire motion of the object, we can focus only on the initial parameters and the final parameters of an object's movement. The law of conservation of energy states that **the total energy of the system is conserved**:

$$\sum E_i = \sum E_f.$$

By knowing the total initial energy of a system, we can find the final energy in the system, and from the final energy we can calculate the final motion parameters.

Units

In math we work with numbers—we solve questions where the answers are numbers without dimensions like 3, 5 or 12.34. The universal power of math comes precisely from this abstraction of things into numbers. We could be solving for the number of sheep in a pen, the surface area of a sphere, or the annual revenue of your startup; we can apply the same mathematical techniques to each example, even though the numbers we use will represent very different kinds of quantities.

In physics we use numbers too, and because physics deals with real-world concepts, each number in physics always comes with a *measurement unit*. We must pay attention to the units of physical quantities and—most importantly—to distinguish between the different *dimensions* of numerical quantities. An answer in physics is a number that represents a length, a time, a velocity, an acceleration, or some other physical quantity. It doesn't make sense to add a *time* and a *mass*, because the two numbers measure different kinds of quantities.

Here's a list of some kinds of quantities discussed in this chapter:

Dimension	**SI unit**	**Other units**	**Measured with**
time	[s]	[h], [min]	clock
length	[m]	[cm], [mm], [ft], [in]	metre tape
velocity	[m/s]	[km/h], [mi/h]	speedometer
acceleration	[m/s^2]		acceleroometer
mass	[kg]	[g], [lb]	scale

Appendix A (see page 452 in the back of the book) provides a more detailed list of the International System of Units (abbreviated SI for *Système International*).

The units of physical quantities are indicated in square brackets throughout the lessons of this chapter. In your equations, you should always try to keep in mind the units for different physics quantities. Sometimes you'll be able to catch yourself making an error because the units will not come out right. If I ask you to calculate the maximum height a ball will reach, I expect your answer to be a length measured in [m] and not some other kind of quantity like a velocity $[\text{m/s}]$ or an acceleration $[\text{m/s}^2]$ or an area $[\text{m}^2]$. An answer in $[\text{ft}]$ would also be acceptable since this is also a length, and it can be converted to metres using $1[\text{ft}] = 0.3048[\text{m}]$ (see page 453 for other conversion ratios). Learn to watch out for the units and dimensions of physical quantities, and you'll have an easy time in physics. They are an excellent error-checking mechanism.

We'll begin our physics journey by starting with the familiar subject of kinematics which we studied in Chapter 2. Now that you know about vectors, we can study two-dimensional kinematics problems, such as the motion of a projectile.

4.2 Projectile motion

Ever since the invention of gun powder, generation after generation of men have thought of countless different ways to hurtle shrapnel at each other. Indeed, mankind has been stuck to the idea of two-dimensional projectile motion like flies on shit. As long as there is money to be made in selling weapons, and so long as the media continues to justify the legitimacy of the use of these weapons, it is likely the trend will continue.

It is therefore imperative for anyone interested in reversing this trend to learn about the physics of projectile motion. You need to know the techniques of the enemy (the industrial military complex) before you can fight them. We'll see that projectile motion is nothing more than a pair of parallel one-dimensional kinematics problems: UVM in the x-direction and UAM in the y-direction.

Concepts

The basic concepts of kinematics in two dimensions are:

- $\hat{x}, \hat{y}$: the xy-coordinate system
- t: time, measured in seconds
- $\vec{r}(t) \equiv (x(t), y(t))$: the position vector of the object at time t
- $\vec{v}(t) \equiv (v_x(t), v_y(t))$: the velocity vector of the object
- $\vec{a}(t) \equiv (a_x(t), a_y(t))$: the acceleration vector of the object

We will use the following terminology when analyzing the motion of an object that starts from an *initial* point and travels to a *final* position:

- $t_i = 0$: the initial time
- t_f: the final time
- $\vec{v}_i = (v_x(0), v_y(0)) = (v_{ix}, v_{iy})$: the initial velocity at $t = 0$
- $\vec{r}_i = (x(0), y(0)) = (x_i, y_i)$: the initial position at $t = 0$
- $\vec{r}_f = \vec{r}(t_f) = (x(t_f), y(t_f)) = (x_f, y_f)$: the position at $t = t_f$

Definitions

Motion in two dimensions

We use the position vector $\vec{r}(t)$ to describe the x and y coordinates of the projectile as a function of time:

$$\vec{r}(t) = (x(t), y(t)).$$

We use x to describe the horizontal distance travelled by the projectile and y to describe the height of the projectile.

The velocity of the projectile is the derivative of its position:

$$\vec{v}(t) = \frac{d}{dt}\left(\vec{r}(t)\right) = \left(\frac{dx(t)}{dt}, \frac{dy(t)}{dt}\right) = (v_x(t), v_y(t)).$$

The initial velocity is an important parameter of the motion:

$$\vec{v}(0) = (v_x(0), v_y(0)) = (v_{ix}, v_{iy}) = (\|\vec{v}_i\|\cos\theta, \|\vec{v}_i\|\sin\theta) = \|\vec{v}_i\|\angle\theta.$$

The initial velocity vector can be expressed as components (v_{ix}, v_{iy}), or in the length-and-direction form $\|\vec{v}_i\|\angle\theta$, where θ measures the angle between $\vec{v}_i$ and the x-axis.

The acceleration of the projectile is

$$\vec{a}(t) = \frac{d}{dt}\left(\vec{v}(t)\right) = (a_x(t), a_y(t)) = (0, -9.81).$$

We know the exact value of the object's acceleration in both the x-direction and the y-direction. There is zero acceleration in the x-direction because there are no horizontal forces acting on the projectile (we ignore the effects of air friction). In the y-direction we have a uniform downward acceleration due to gravity.

Projectile motion

The motion of a projectile can be described by two sets of equations.

In the x-direction, the motion is described by the uniform velocity motion (UVM) equations of motion:

$$\begin{aligned} x(t) &= v_{ix}t + x_i, \\ v_x(t) &= v_{ix}. \end{aligned}$$

We use the UVM equations of motion for $x(t)$ and $v_x(t)$ because there are no horizontal forces acting on the object, and by extension the object experiences zero acceleration in the x-direction: $a_x = 0$.

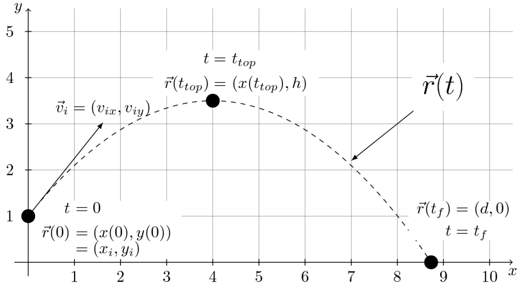
In the y-direction, the constant, downward pull of gravity produces uniformly accelerated motion (UAM). The equations of motion in the y-direction are

$$\begin{aligned} y(t) &= \tfrac{1}{2}(-9.81)t^2 + v_{iy}t + y_i, \\ v_y(t) &= (-9.81)t + v_{iy}, \\ v_{yf}^2 &= v_{iy}^2 + 2(-9.81)(\Delta y). \end{aligned}$$

The equations in the y-direction correspond to the standard UAM equations with $a = -9.81$[m/s^2].

Example

In this example, we'll analyze all aspects of the motion of a projectile. An object is thrown from an initial height of 1[m], at an initial velocity of 8.96[m/s], at an angle of $51.3°$ to the ground. Calculate the maximum height h the object will reach, and the distance d where the object will hit the ground.



Our first step when reading any physics problem is to extract all quantitative information from the problem statement. The object's initial position is $\vec{r}(0) = (x_i, y_i) = (0, 1)$[m]. Its initial velocity is $\vec{v}_i = 8.96\angle 51.3°$[m/s], which is $\vec{v}_i = (8.96\cos 51.3°, 8.96\sin 51.3°) = (5.6, 7)$[m/s] in component form.

Next, plug the values of $\vec{r}_i = (0, 1)$[m] and $\vec{v}_i = (5.6, 7)$[m/s] into the equations of motion for the x and y directions:

$$\begin{aligned} x(t) &= 5.6t + 0, & y(t) &= \tfrac{1}{2}(-9.81)t^2 + 7t + 1, \\ v_x(t) &= 5.6, & v_y(t) &= (-9.81)t + 7. \end{aligned}$$

When the object reaches its maximum height, it will have zero velocity in the y-direction: $v_y(t_{\text{top}}) = 0$. We can use this fact along with the $v_y(t)$ equation to find $t_{\text{top}} = 7/9.81 = 0.714$[s]. The maximum height is then obtained by evaluating the function $y(t)$ at $t = t_{\text{top}}$. We obtain $h = y(t_{\text{top}}) = \frac{1}{2}(-9.81)(0.714)^2 + 7(0.714) + 1 = 3.5$[m].

To find d, we must find the time t_f when the object hits the ground. We can find t_f by solving the quadratic equation $0 = y(t_f) = \frac{1}{2}(-9.81)(t_f)^2 + 7(t_f) + 1$. The solution is $t_f = 1.55$[s]. We then plug this time value into the equation for $x(t)$ to find $d = x(t_f) = 5.6(1.55) + 0 = 8.68$[m]. Can you verify that these answers match the trajectory illustrated in the figure?

Explanations

Coordinate system

Before you begin solving any projectile motion problem, you should make a diagram of what is going on. In your diagram, be sure to clearly indicate the coordinate system with respect to which you'll measure x, y, v_x, and v_y. The values you plug into the equations of motion are measured with respect to this coordinate system; for example, a velocity v_x in the opposite direction of the x-axis is represented as a negative number.

Uniform velocity motion in the x-direction

Ignoring the effects of air friction means there are no forces and no acceleration in the x-direction, so $a_x = 0$. As a consequence, the velocity will be constant. Constant velocity means the projectile will keep whatever x-velocity you give it when you throw it. Therefore the UVM equations describe the projectile's motion in the x-direction:

$$\begin{aligned} a_x(t) &= 0\,, \\ v_x(t) &= v_{ix}\,, \\ x(t) &= v_{ix}t + x_i\,. \end{aligned}$$

Uniform acceleration motion in the y-direction

The pull of gravity acts in the negative y-direction. Gravity is a constant downward acceleration equal to $g = 9.81$[m/s^2]. The motion in the y-direction is therefore described by the UAM equations with acceleration $a_y = -g = -9.81$[m/s^2]:

$$\begin{aligned} a_y(t) &= -g\,, \\ v_y(t) &= (-g)t + v_{iy}\,, \\ y(t) &= \tfrac{1}{2}(-g)t^2 + v_{iy}t + y_i\,. \end{aligned}$$

We can also employ the fourth equation of motion

$$v_{fy}^2 = v_{iy}^2 + 2(-g)(\Delta y),$$

to relate the object's final velocity to its initial velocity and its displacement in the y-direction, $\Delta y = y_f - y_i$.

Examples

Roach throw

You are sitting comfortably on a bench in the park and you have a small piece of garbage in your hand. Not far from you is a garbage bin. Since you're feeling lazy and relaxed, you can't be bothered to walk to the bin and dispose of said garbage particle, so you decide to throw it into the bin from where you are sitting. Let's call the garbage particle r for short. Imagine a coordinate system centred below your feet. The point $(0, 0)$ is where you are sitting, and the point $(x = 0, y = 1.4)$[m] is the initial position of the particle r as you prepare to throw it.

Suppose the distance to the garbage bin is 3 metres and the bin is 1 metre tall. Can you calculate the initial velocity $\vec{v}_i$ the particle r needs in order to land in the garbage bin? Assume you flick the particle from your fingers so it flies straight along the x-axis; in other words, you do not give the particle any initial y velocity so $v_{iy} = 0$.

To describe the motion of r, all you need to know is the initial position $\vec{r}(0) = (x(0), y(0))$, and the initial velocity $\vec{v}_i = \vec{v}(0) = (v_x(0), v_y(0))$. You can then plug these values into the projectile equations of motion:

$$\begin{aligned} x(t) &= v_{ix}t + x_i, \\ y(t) &= \tfrac{1}{2}a_y t^2 + v_{iy}t + y_i. \end{aligned}$$

Most physics word problems will follow this pattern. The problem statement gives you some information about the initial conditions and the desired final conditions, and asks you to solve for the *unknown*—the one variable that is not given in the problem statement.

Can you carry out the necessary calculations in this case? I don't mean to stress you out, but sitting next to you is your 110[kg] pure-muscle Chilean friend who has two kids and *really* gets pissed off at people who throw garbage around in the park. You don't want to piss him off so you better get that initial velocity right!

Okay, from here we can switch into high gear because we have everything set up nicely. We know the general equations of motion for UVM in x and UAM in y are

$$\begin{aligned} x(t) &= v_{ix}t + x_i, \\ y(t) &= \tfrac{1}{2}a_y t^2 + v_{iy}t + y_i\,. \end{aligned}$$

More specifically, we know the y acceleration is due to gravity, so we have

$$\begin{aligned} x(t) &= v_{ix}t + x_i, \\ y(t) &= \tfrac{1}{2}(-9.81)t^2 + v_{iy}t + y_i\,. \end{aligned}$$

We also know the position at $t = 0$ is $(x_i, y_i) = (0, 1.4)$[m], and that at some $t_f > 0$ the particle will be flying into the bin at $(x(t_f), y(t_f)) = (3, 1)$[m].

Substituting all the known quantities into the general equations, we obtain

$$\begin{aligned} x(t_f) &= 3 = v_{ix}t_f + 0, \\ y(t_f) &= 1 = \tfrac{1}{2}(-9.81)t_f^2 + v_{iy}t_f + 1.4\,. \end{aligned}$$

Furthermore, as the problem specifies, we can assume the initial velocity of the projectile is purely horizontal ($v_{iy} = 0$). Thus, we must solve the pair of equations,

$$\begin{aligned} 3 &= v_{ix}t_f, \\ 1 &= 1.4 - 4.9t_f^2\,, \end{aligned}$$

where v_{ix} and t_f are the two unknowns.

From this step, it should be clear where the story is going. First we solve for t_f in the second equation:

$$t_f = \sqrt{\frac{(1-1.4)}{-4.9}} = \sqrt{\frac{-0.4}{-4.9}} = \sqrt{4/49} = 2/7 \approx 0.2857\text{[s]}.$$

We can now solve for v_{ix} in the first equation:

$$v_{ix} = \frac{3}{t_f} = \frac{3 \cdot 7}{2} = \frac{21}{2} = 10.5[\text{m/s}].$$

You flick r with you finger at an initial velocity of $\vec{v}_i = (10.5, 0)$[m/s] and the particle flies straight into the garbage bin. Success!

Freedom and democracy

An American F-18 is flying above Iraq. It is carrying two bombs. One bomb is named "freedom" and weighs 200[kg]; the other is called "democracy" and packs a mass of 500[kg]. If the plane is flying horizontally with speed $v_i = 300$[m/s] and drops both bombs from a height of 2000[m], how far will the bombs travel before they hit the ground? Which city will get freedom and which city will get democracy?

The equations of motion for the bombs are

$$x(t) = v_{ix}t + x_i = 300t + 0$$

and

$$y(t) = \tfrac{1}{2}(-9.81)t^2 + v_{iy}t + y_i = -4.9t^2 + 2000.$$

To find where the bombs will land, the first step is to calculate the time of flight. We solve for t_f in the equation $y(t_f) = 0$ and find $t_f = 20.20$[s]. We can then find the final x-position where the bombs hit the ground from the first equation: $x_f = x(20.20) = 6060$[m]. Both bombs hit the same town, the one located 6.06[km] from the launch point. Observe that the bombs' masses did not play any part in the final equations of motion.

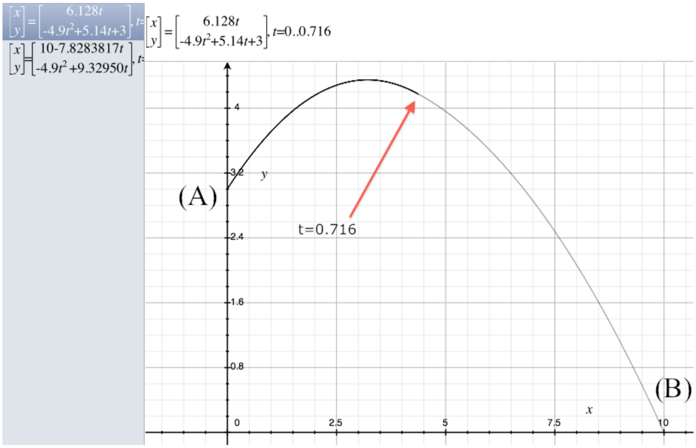
Let's be real. The scenario at hand is essentially what the people in Washington are talking about when they say they are bringing freedom and democracy to the Middle East. A monstrous amalgamation of warmongering corporations, weak politicians, and special-interest lobby groups make a complete mockery of the political process. In order to see an end to world conflict, I think the entire military-industrial complex needs to be dismantled. How can we stop them, you ask? In my opinion, the best way to fight the System is not to work for the System. If some recruiters from that sector comes to offer you a job one day because you're a math expert, tell them to scram.

Interception

With all those people launching explosive projectiles at each other, a need develops for *interception* systems that can throw counter-projectiles at the incoming projectiles and knock them out of the air.

Let's see how we can intercept an incoming ball (A) launched from $\vec{r}_{Ai} = (0, 3)$ with initial velocity $\vec{v}_{Ai} = (8\cos(40), 8\sin(40))$. As an interception device, you have at your disposal a ball launcher placed at $\vec{r}_{Bi} = (10, 0)$ with a fixed firing angle of $50°$. You position the launcher so it faces the incoming ball.

The launcher has a variable launch speed w[m/s], which you can choose. You want to fire an intercepting ball, which will have an initial velocity $\vec{v}_{Bi} = (-w\cos(50), w\sin(50))$, so it intercepts the ball (A) in midair. What initial velocity w is required for the balls to hit each other? At which time t will the collision occur?



$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6.128t \\ -4.9t^2+5.14t+3 \end{bmatrix}, t=$

$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10-7.8283817t \\ -4.9t^2+9.32950t \end{bmatrix}, t=$

$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6.128t \\ -4.9t^2+5.14t+3 \end{bmatrix}, t=0..0.716$

Figure 4.2: Midair interception. The incoming ball (A) (black line) is intercepted by the ball (B) (grey line).

As far as kinematics is concerned, this is a standard projectile motion problem **times two**. You have ball (A), which has the equations of motion,

$$\begin{aligned} x_A(t) &= v_{Aix}t + x_{Ai} = 8\cos(40)t + 0, \\ y_A(t) &= \tfrac{1}{2}(-9.81)t^2 + v_{Aiy}t + y_{Ai} = -4.9t^2 + 8\sin(40)t + 3. \end{aligned}$$

You also have ball (B), which has the equations of motion,

$$\begin{aligned} x_B(t) &= v_{Bix}t + x_{Bi} = -w\cos(50)t + 10, \\ y_B(t) &= \tfrac{1}{2}(-9.81)t^2 + v_{Biy}t + y_{Bi} = -4.9t^2 + w\sin(50)t + 0. \end{aligned}$$

We want the balls to collide, so at some point they will have the same coordinates $\vec{r}_A = \vec{r}_B$, which is another way of saying

$$(x_A(t), y_A(t)) = (x_B(t), y_B(t)).$$

The x-coordinates must match, and the y-coordinates must match. We express these conditions through the following two equations:

$$\begin{aligned} 8\cos(40)t + 0 &= -w\cos(50)t + 10, \\ -4.9t^2 + 8\sin(40)t + 3 &= -4.9t^2 + w\sin(50)t + 0, \end{aligned}$$

which we must solve simultaneously.

To solve, we cancel $-4.9t^2$ on both sides of the bottom equation:

$$\begin{aligned} 8\cos(40)t &= -w\cos(50)t + 10, \\ 8\sin(40)t + 3 &= w\sin(50)t. \end{aligned}$$

This is a set of two equations with two unknowns, and we can solve it. It's not going to be easy, since we can't cleanly isolate t or w using standard substitution techniques. However, there is a trick! We can divide the two equations. If $A = B$ and $C = D \neq 0$ then $A/C = B/D$, and this is what we'll use. To prepare for this step, let's rearrange the equations a bit so all the w-containing terms stand alone on the right side:

$$\begin{aligned}10 - 8\cos(40)t &= w\cos(50)t,\\ 8\sin(40)t + 3 &= w\sin(50)t.\end{aligned}$$

We'll now divide the bottom equation by the top equation to obtain

$$\frac{8\sin(40)t + 3}{10 - 8\cos(40)t} = \frac{w\sin(50)t}{w\cos(50)t} = \tan(50).$$

Rearranging the expression, we find

$$8\sin(40)t + 3 = \tan(50)(10 - 8\cos(40)t).$$

Collect all the t terms to one side to obtain

$$[8\sin(40) + 8\cos(40)\tan(50)]t = 10\tan(50) - 3,$$

and finally

$$t = \frac{10\tan(50) - 3}{8\sin(40) + 8\cos(40)\tan(50)} = 0.7165[\text{s}].$$

We can now plug values into any of the above equations to find the value of w. For example, try plugging the value of $t = 0.7165$[s] into

$$10 - 8\cos(40)t = w\cos(50)t$$

to find

$$10 - 8\cos(40)(0.7165) = w\cos(50)(0.7165),$$

which leads to $w = \frac{10-8\cos(40)(0.7165)}{\cos(50)(0.7165)} = 12.1788$ [m/s].

Let's check this answer. If we substitute an initial velocity $w = 12.1788$[m/s] into the equations of motion and plot the two trajectories on the computer we obtain the graph shown in Figure 4.2. As you can see, the trajectories intersect at time $t = 0.7165$[s] as expected.

Discussion

I want to point out that you need no new physics information to understand the motion of projectiles. Projectile motion is a two-dimensional kinematics problem that can be broken down into two parts: the x-direction (described by the UVM equations) and the y-direction (described by the UAM equations).

Links

[Eisenhower on the danger posed by the military-industrial complex]

> *"Only an alert and knowledgeable citizenry can compel the proper meshing of the huge industrial and military machinery of defence with our peaceful methods and goals."*

http://www.youtube.com/watch?v=8y06NSBBRtY

4.3 Forces

Like a shepherd who brings back stray sheep, we need to rescue the word *force* and give it precise meaning. In physics, force means something very specific. I'm not talking about "the force" from Star Wars, nor the "force of public opinion," nor the "*force* in the battle of good versus evil."

In physics, force refers to an amount of push or pull exerted on an object. Forces are vector quantities measured in Newtons [N]. In this section, we'll explore all the different kinds of forces.

Concepts

- $\vec{F}$: a force. This is something the object "feels" as a pull or a push. Forces are vector quantities, so you must always keep in mind the direction in which they act.

- $k, G, m, \mu_s, \mu_k, \ldots$: parameters on which the force F may depend. For example, the heavier an object is (the larger the m parameter), the larger its gravitational pull will be. This relationship is expressed by the equation $\vec{W} = -9.81m\hat{\jmath}$, where $\hat{\jmath}$ points toward the sky.

Kinds of forces

Next, we'll review all the forces you're supposed to know for a standard mechanics class, and define the relevant parameters for each kind of force. You need to practice exercises using each of these forces, until you start to *feel* how they act.

Gravitation

The force of gravity exists between any two massive objects. The magnitude of the gravitational force between two objects of mass M[kg] and m[kg] separated by a distance r[m] is given by the formula

$$F_g = \frac{GMm}{r^2},$$

where $G = 6.67 \times 10^{-11}[\frac{\text{Nm}^2}{\text{kg}^2}]$ is the *gravitational constant*. This is one of Newton's biggest discoveries—the famous one-over-r-squared law.

On the surface of the Earth, which has mass $M = 5.972 \times 10^{24}$[kg] and radius $r = 6.367 \times 10^{6}$[m], the force of gravity on an object of mass m is given by

$$F_g = \frac{GMm}{r^2} = \underbrace{\frac{GM}{r^2}}_{g} m = 9.81m = W.$$

We call this force the *weight* of the object, and to be precise we should write $\vec{W} = -mg\hat{\jmath}$ to indicate that the force acts *downward* in the negative y-direction. Verify using your calculator that $\frac{GM}{r^2} = 9.81 \equiv g$.

Force of a spring

A spring is a piece of metal twisted into a coil that has a certain natural length. The spring will resist any attempts to stretch or compress it. The force exerted by a spring is given by

$$\vec{F}_s = -k\vec{x},$$

where x is the amount by which the spring is displaced from its natural length, and the constant k[N/m] is a measure of the spring's *strength*. Note the negative sign indicates that the spring always acts to oppose the displacement. If you try to stretch the spring, displacing it in the positive x-direction, then the force of the spring will pull against you (the spring will pull in the negative x-direction). Similarly, if you try to compress the spring (a displacement in the negative x-direction), the spring will push back against you, in the positive x-direction.

Normal force

The *normal* force is the force between two surfaces in contact. In this context, the word *normal* means "perpendicular to the surface of." The reason my coffee mug is not falling to the floor right now is that the table exerts a normal force $\vec{N}$ on the mug, keeping it in place.

Force of friction

In addition to the normal force between surfaces, there is also the force of friction $\vec{F}_f$, which acts to impede any sliding motion between the surfaces. There are two kinds of friction forces, and both are proportional to the amount of normal force between the surfaces:

$$\max\{\vec{F}_{fs}\} = \mu_s\|\vec{N}\| \ \text{(static)}, \quad \text{and} \quad \vec{F}_{fk} = \mu_k\|\vec{N}\| \ \text{(kinetic)},$$

where μ_s and μ_k are the static and dynamic *friction coefficients*. It makes sense that the force of friction should be proportional to the magnitude of the normal force $\|\vec{N}\|$, since the harder the two surfaces push against each other, the more difficult it becomes to make them slide. The above equations give mathematical precision to this intuitive logic.

The static force of friction acts on objects that are not moving. It describes the *maximum* amount of friction that can exist between two objects. If a horizontal force greater than $F_{fs} = \mu_s N$ is applied to the object, then it will start to slip. The kinetic force of friction acts when two objects are sliding relative to each other. It always acts in the direction opposite to the motion.

Tension

A force can also be exerted on an object remotely by attaching a rope to the object, and pulling the rope. The force exerted on the object will be equal to the rope's *tension* $\vec{T}$. Note that tension always pulls *away* from an object: you can pull but you can't push a dog by its leash.

Discussion

Viewing the interactions between objects in terms of the forces that act between them gives us a powerful tool for thinking and analyzing physics problems. The following section shows you how to draw force diagrams that account for all the forces acting on an object.

4.4 Force diagrams

Welcome to Force-Accounting 101. Here, we'll learn how to identify all the forces acting on an object and use Newton's 2nd law $\sum \vec{F} = \vec{F}_{\text{net}} = m\vec{a}$ to predict the resulting acceleration of the object.

Concepts

Newton's second law describes a relationship between three quantities:

- m: the *mass* of an object
- $\vec{F}_{\text{net}}$: the net force acting on the object
- $\vec{a}$: the *acceleration* of the object

Forces and accelerations are vectors. To work with vectors, we work with their *components*:

- F_x: the *component* of $\vec{F}$ in the x-direction
- F_y: the *component* of $\vec{F}$ in the y-direction

Vectors are meaningless unless it is clearly explained which *coordinate system* they are expressed in respect to.

- The x-axis: The x-axis is usually horizontal and points to the right. For problems with inclines, it will be more convenient to use an inclined x-axis that is parallel to the slope.
- The y-axis: The y-axis is always *perpendicular* to the x-axis.
- The $\hat{\imath}$ and $\hat{\jmath}$ vectors: These are unit vectors in the x and y directions, respectively. Any vector can be written as $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath}$ or as $\vec{v} = (v_x, v_y)$.

We can write any force vector in three equivalent ways:

$$\vec{F} \equiv F_x\hat{\imath} + F_y\hat{\jmath} \equiv (F_x, F_y) \equiv \|\vec{F}\|\angle\theta.$$

What types of forces are represented in force diagrams?

- $\vec{W} \equiv \vec{F}_g$: the force of gravity. The *weight* of an object is the force the object feels due to gravity. The gravitational pull always points downward, toward the centre of the Earth.
- $\vec{T}$: the tension in a rope. Tension always pulls *away* from the object.
- $\vec{N}$: the normal force. The normal force is part of the contact force between two surfaces.
- $\vec{F}_{fs} = \mu_s\|\vec{N}\|$: the static force of friction
- $\vec{F}_{fk} = \mu_k\|\vec{N}\|$: the kinetic force of friction
- $\vec{F}_s = -kx$: the force (pull or push) of a spring that is displaced (stretched or compressed) by x metres

Formulas

Newton's 2nd law

The sum of the forces acting on an object, divided by the object's mass, gives the acceleration of the object:

$$\sum \vec{F} \equiv \vec{F}_{\text{net}} = m\vec{a}. \tag{4.1}$$

Vector components

If a vector $\vec{v}$ makes an angle θ with the x-axis, then

$$v_x = \|\vec{v}\| \cos\theta \qquad \text{and} \qquad v_y = \|\vec{v}\| \sin\theta.$$

The vector $v_x\hat{\imath}$ corresponds to the part of $\vec{v}$ that points in the x-direction.

Shortly, I'll be asking you over and over again to

find the component of $\vec{F}$ in the ? direction,

which is another way of asking you to find the number $v_?$.

The answer is usually equal to the length $\|\vec{F}\|$ multiplied by either $\cos$ or $\sin$ or sometimes -1, **depending on way the coordinate system is chosen**. So don't guess. Look at the coordinate system. If the vector points in the direction where x increases, then v_x should be a positive number. If $\vec{v}$ points in the opposite direction, then v_x should be negative.

To add forces $\vec{F}_1$ and $\vec{F}_2$, you need to add their components:

$$\vec{F}_1 + \vec{F}_2 = (F_{1x}, F_{1y}) + (F_{1x}, F_{2y}) = (F_{1x} + F_{2x}, F_{1y} + F_{2y}) = \vec{F}_{\text{net}}.$$

However, instead of dealing with vectors in the bracket notation, when solving force diagrams it is easier to simply write the x equation on one line, and the y equation on a separate line below it:

$$F_{\text{net},x} = F_{1x} + F_{2x},$$

$$F_{\text{net},y} = F_{1y} + F_{2y}.$$

It's a good idea to always write those two equations together as a block, so it's clear the first row represents the x dimension and the second row represents the y dimension for the same problem.

Force check

It is important to account for *all* the forces acting on an object. First of all, any object with mass on the surface of the Earth will feel a downward gravitational pull of magnitude $F_g = W = m\vec{g}$. Then you must consider whether any of the other forces are present: $\vec{T}$, $\vec{N}$, $\vec{F}_f$, and $\vec{F}_s$. Any time you see a rope tugging on the object, you can know there must be some tension $\vec{T}$, which is a force vector pulling on the object. Any time you have an object sitting on a surface, the surface pushes back with a *normal* force $\vec{N}$. If the object slides on the surface, then a force of friction acts against the direction of motion:

$$F_{fk} = \mu_k \|\vec{N}\|.$$

If the object is not moving, you must use the equation for the static force of friction. The maximum static friction force that the contact between the object and the ground can support before the object starts to slip is

$$\max\{F_{fs}\} = \mu_s \|\vec{N}\|.$$

If you see a spring that is either stretched or compressed by the object, then you must account for the spring force. The force of a spring is *restorative*; it always acts against the deformation you exert on the spring. If you stretch the spring by x[m], it will try to pull itself back to its normal length with a force of

$$\vec{F}_s = -kx\hat{\imath}.$$

The constant of proportionality k is called the *spring constant* and is measured in [N/m].

Solving force diagrams

We'll now explain how to solve dynamics problems. We'll first describe the general procedure in terms of a sequence of steps. Afterward, we'll illustrate how to use this procedure through a series of examples.

The steps for solving dynamics problems are as follows:

1. Draw a force diagram focused on the object, and indicate all the forces acting on the object.
2. Choose a coordinate system, and indicate clearly in the diagram what you are calling the positive x-direction, and what you are calling the positive y-direction. All quantities in the subsequent equations will be expressed *with respect to* this coordinate system.
3. Write the $\vec{F} = m\vec{a}$ template:

$$\begin{aligned}\sum F_x &= \qquad\qquad\qquad = ma_x,\\ \sum F_y &= \qquad\qquad\qquad = ma_y.\end{aligned}$$

4. Fill in the template by calculating the x and y components of each force acting on the object ($\vec{W}$, $\vec{N}$, $\vec{T}$, $\vec{F}_{fs}$, $\vec{F}_{fk}$, $\vec{F}_s$, as applicable).
5. Solve the equations for the unknown quantities.

I highly recommend you perform some consistency checks after Step 4 by checking the sign of each force: if a force in the diagram is acting in the x-direction, then its component must be positive. If the force is acting in the direction opposite to the x-axis, then its component must be negative. You should also check to make sure that whenever $F_x = \|\vec{F}\|\cos\theta$, then $F_y = \|\vec{F}\|\sin\theta$. If, instead of θ, you use the angle ϕ defined with respect to the y-axis, then the roles of $\sin$ and $\cos$ will change: $F_x = \|\vec{F}\|\sin\phi$ and $F_y = \|\vec{F}\|\cos\phi$.

Examples

Block on a table

You place a block of mass m on the table. Since the block has mass m, its weight $\vec{W}$ pulls down on it, yet the table stops the block from dropping to the floor. The table pushes back on the block with a normal force $\vec{N}$.

Steps 1 and 2: Draw the force diagram and choose a coordinate system:

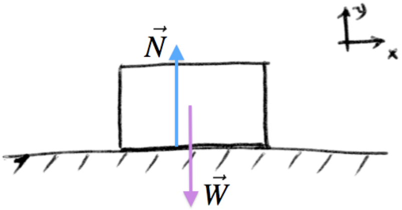


Figure 4.3: A block sitting on a table. The weight of the block $\vec{W}$ is counteracted by the normal force $\vec{N}$.

Step 3: Next, write the empty force diagram equations template:

$$\sum F_x = \qquad\qquad = ma_x,$$
$$\sum F_y = \qquad\qquad = ma_y.$$

Step 4: There are no forces acting in the x-direction, and the block is not moving, so $a_x = 0$. In the y-direction, we have the force of gravity and the normal force exerted by the table:

$$\sum F_x = 0 = 0,$$
$$\sum F_y = N - mg = 0.$$

We set $a_y = 0$, as we can *see* that the block is just sitting there on the table without moving. The technical term for situations where $a_x = 0, a_y = 0$ is called *static equilibrium*. Force diagrams with static equilibrium are easy to solve because the entire right-hand side is equal to zero, which means the forces acting on the object must be counter-balancing each other.

Step 5: Suppose the teacher now asks, "What is the magnitude of the normal force?" By looking at the second equation, you can answer, "$N = mg$ bro!"

Moving the fridge

You are trying to push your fridge across the kitchen floor. It weighs quite a lot, and is strongly "gripping" the floor when you try to push it. The static coefficient of friction between the bottom of your fridge and the tiles on the floor is μ_s. How much force $\vec{F}_{\text{ext}}$ will it take to cause the fridge to start moving?

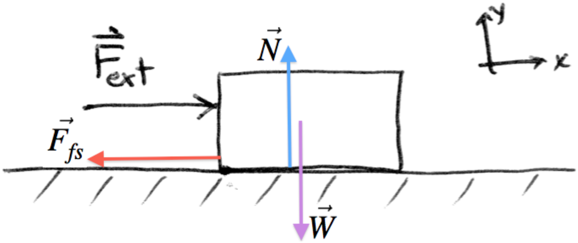


Figure 4.4: How strongly do you need to push before the fridge starts to slip?

$$\sum F_x = F_{\text{ext}} - F_{fs} = 0,$$
$$\sum F_y = N - mg = 0.$$

If you push with a force of $F_{\text{ext}} = 30$[N], the fridge will push back via its connection to the floor with a force $F_{fs} = 30$[N]. If you push harder, the fridge will push back harder and it still will not move. Only when you reach the slipping threshold will the fridge move. This means you'll need to push with a force equal to the *maximum* static friction force $F_{fs} = \mu_s N$, so we have

$$\sum F_x = F_{\text{ext}} - \mu_s N = 0,$$
$$\sum F_y = N - mg = 0.$$

To solve for F_{ext}, first look at the bottom equation and isolate $N = mg$, then substitute the value of N in the top equation to find $F_{\text{ext}} = \mu_s mg$.

Friction slowing you down

Okay, now you're moving the fridge at a steady pace across the room:

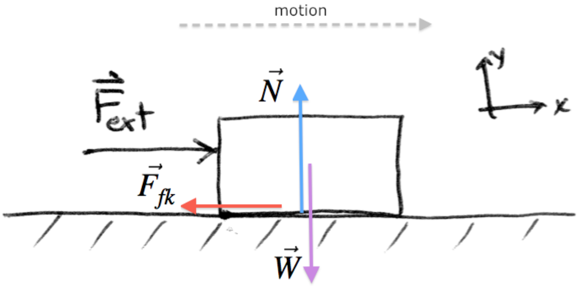


Figure 4.5: The fridge is moving. What is the magnitude of the external force $\vec{F}_{\text{ext}}$ required to counterbalance the kinetic force of friction?

Your equation of motion is expressed as

$$\sum F_x = F_{\text{ext}} - F_{fk} = ma_x,$$
$$\sum F_y = N - mg = 0.$$

In particular, if you want to keep a steady speed ($v = \text{const}$) as you move the fridge across the room, you'll need to push the fridge with a force that exactly balances the friction force and keeps $a_x = 0$.

To find the value of F_{ext} that allows you to keep a constant speed, solve

$$\sum F_x = F_{\text{ext}} - \mu_k N = 0,$$
$$\sum F_y = N - mg = 0.$$

The above set of equations are similar to the equations we obtained for the fridge that was not moving. The only difference is the kinetic coefficient of friction μ_k replaces the static coefficient of friction μ_s. Keeping the fridge moving with a constant velocity requires an external force $F_{\text{ext}} = \mu_k mg$. Generally, $\mu_k < \mu_s$, so less force is needed to keep the fridge moving than is needed to start the fridge moving.

Let's approach this whole friction thing from a different slant.

Incline

At this point, my dear readers, we're delving into the crucial question that you will—without a doubt—be asked to solve in your homework or at the final exam.

A block is sliding down an incline. What is its acceleration?

Step 1: Draw a diagram that includes the block's weight $\vec{W}$, the normal force $\vec{N}$, and the friction force $\vec{F}_{fk}$.

Step 2: Choose the coordinate system to be tilted along the incline. This is important because, in this coordinate system, the block's motion is purely in the x-direction, while the y-direction remains static.

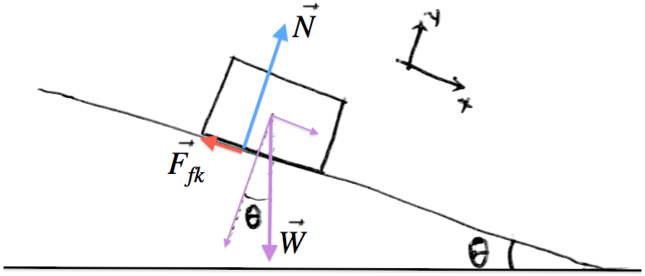


Figure 4.6: A block sliding down an incline with angle θ. What is the block's acceleration?

Steps 3 and 4: Let's copy the empty template and fill in the equations:

$$\sum F_x = \|\vec{W}\| \sin\theta - F_{fk} = ma_x,$$
$$\sum F_y = N - \|\vec{W}\| \cos\theta \quad = 0;$$

or, substituting the values that we know,

$$\sum F_x = mg\sin\theta - \mu_k N = ma_x,$$
$$\sum F_y = N - mg\cos\theta \quad = 0.$$

Step 5: From the y equation, we obtain $N = mg\cos\theta$, which we substitute into the x equation to obtain

$$a_x = \frac{1}{m}\left(mg\sin\theta - \mu_k mg\cos\theta\right) = g\sin\theta - \mu_k g\cos\theta.$$

Bathroom scale

You have a spring in your bathroom scale with spring constant k, on which you place a block of mass m. By what length Δy will the spring be compressed?

Step 1, 2: Draw a before and after picture with the y-axis placed at the *natural* length of the spring.

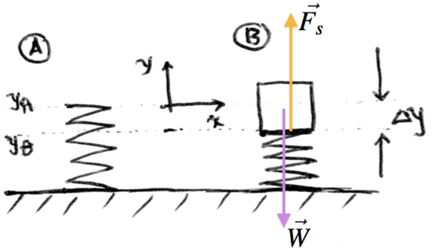


Figure 4.7: A bathroom scale is compressed by a distance Δy when an object of mass m is placed on it.

Steps 3 and 4: Filling in the template, we find

$$\sum F_x = 0 = 0,$$
$$\sum F_y = F_s - mg = 0.$$

Step 5: We know the force exerted by a spring is proportional to its displacement according to

$$F_s = -ky_B,$$

so we can find $y_B = -\frac{mg}{k}$. The length of compression is therefore

$$|\Delta y| = \frac{mg}{k}.$$

Two blocks

Now you're ready for a more involved example with two blocks. One block is sitting on a surface, and the other block is falling straight down. The two blocks are connected by a rope. What is the acceleration of the *system* as a whole?

Steps 1 and 2: We have two objects, so we need to draw two force diagrams.

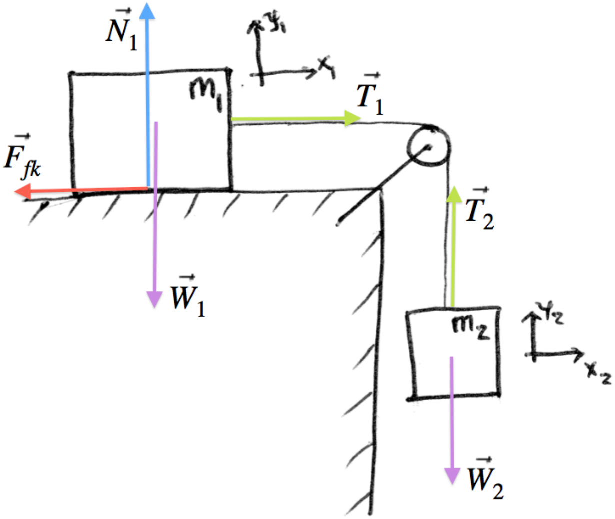


Figure 4.8: A block of mass m_1 is dragged along horizontally by a second block of mass m_2 which is falling vertically. What is the acceleration of the system?

Step 3: We need two sets of equations; one set for the block on the horizontal surface, and one set for the falling block:

$$\sum F_{1x} = \qquad = m_1 a_{x_1} \qquad\qquad \sum F_{2x} = \qquad = m_2 a_{x_2}$$
$$\sum F_{1y} = \qquad = m_1 a_{y_1} \qquad\qquad \sum F_{2y} = \qquad = m_2 a_{y_2}$$

Step 4: We fill in the equations with all the forces drawn in the diagram:

$$\sum F_{1x} = -F_{fk} + T_1 = m_1 a_{x_1} \qquad\qquad \sum F_{2x} = 0 = 0$$
$$\sum F_{1y} = N_1 - W_1 = 0 \qquad\qquad \sum F_{2y} = -W_2 + T_2 = m_2 a_{y_2}$$

Step 5: What connections exist between the two blocks? Since the blocks are connected by the rope, the tension in the rope is equal on both ends, and $T_1 = T_2 = T$. Also, since the rope is a fixed length, the x_1 and y_2 coordinates are related by a constant (though they point in different directions), so it must be that $a_{x_1} = -a_{y_2} = a$.

We'll rewrite the equations in terms of the new *common* variables T and a:

$$\begin{aligned} \sum F_{1x} &= -\mu_k N_1 + T = m_1 a & \qquad \sum F_{2x} &= 0 = 0 \\ \sum F_{1y} &= N_1 - m_1 g = 0 & \qquad \sum F_{2y} &= -m_2 g + T = -m_2 a \end{aligned}$$

Isolate N_1 on the bottom left, and isolate T on the bottom right:

$$\begin{aligned} &\sum F_{1x} = -\mu_k N_1 + T = m_1 a & \qquad &\sum F_{2x} = 0 = 0 \\ &N_1 = m_1 g & \qquad &T = -m_2 a + m_2 g \end{aligned}$$

Substitute the values of N_1 and T into the top left equation:

$$\sum F_{1x} = -\mu_k (m_1 g) + (-m_2 a + m_2 g) = m_1 a.$$

Moving all terms containing a to the right-hand side gives

$$-\mu_k m_1 g + m_2 g = m_1 a + m_2 a = (m_1 + m_2) a.$$

This makes sense if you think about it: two blocks attached with a rope form a single system of collective mass $(m_1 + m_2)$ with two external forces acting on it. From this point of view, the tension T is an *internal* force of the system, so it does not appear in the external force equation.

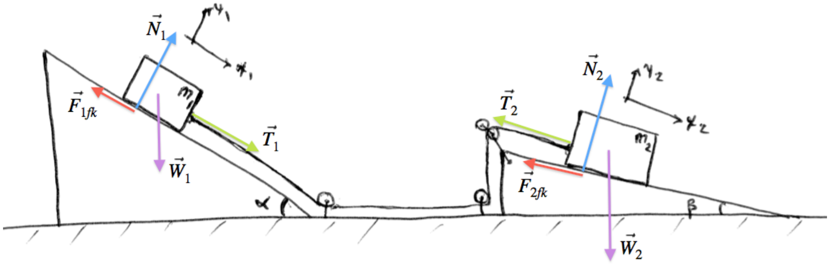
The acceleration of the whole two-block system is

$$a = \frac{m_2 g - \mu_k m_1 g}{m_1 + m_2}.$$

Two inclines

Two inclines? Things just got crazy! We have two inclines, two blocks, a rope, and friction everywhere. As usual, we want to find the acceleration of the system.

Steps 1 and 2: Draw a force diagram with two different coordinate systems, each system adapted for the angle of the incline:



Steps 3 and 4: Make two copies of the template, fill in known forces, and set $a_{y_1} = 0$ and $a_{y_2} = 0$:

$$\begin{aligned}
\sum F_{1x} &= W_1 \sin\alpha - F_{1fk} + T_1 = m_1 a_{x_1}, \\
\sum F_{1y} &= -W_1 \cos\alpha + N_1 \qquad = 0,
\end{aligned}$$

$$\begin{aligned}
\sum F_{2x} &= W_2 \sin\beta - F_{2fk} - T_2 = m_2 a_{x_2}, \\
\sum F_{2y} &= -W_2 \cos\beta + N_2 \qquad = 0.
\end{aligned}$$

Before we continue with this problem, we must identify the connections between the two sets of equations. The tension in the rope is the same, which we will call $T = T_1 = T_2$. Also the two blocks must have the same acceleration since the blocks are moving together $a = a_{x_1} = a_{x_2}$.

Step 5: Rewriting the expression in terms of the common variables T and a, we obtain the new sets of equations:

$$\begin{aligned}
&\sum F_{1x} = m_1 g \sin\alpha - \mu_k N_1 + T = m_1 a, \\
&N_1 = m_1 g \cos\alpha,
\end{aligned}$$

$$\begin{aligned}
&\sum F_{2x} = m_2 g \sin\beta - \mu_k N_2 - T = m_2 a, \\
&N_2 = m_2 g \cos\beta.
\end{aligned}$$

Substitute the values of N_1 and N_2 into the x equations:

$$\sum F_{1x} = m_1 g \sin\alpha - \mu_k m_1 g \cos\alpha + T = m_1 a,$$

$$\sum F_{2x} = m_2 g \sin\beta - \mu_k m_2 g \cos\beta - T = m_2 a.$$

There are many ways to solve for the two unknowns in this pair of equations. We can isolate T in one of the equations, then substitute the value of T into the second equation. Another option is to isolate a in both equations, then set the equations equal to each other.

We'll use the first approach and isolate T in the bottom equation:

$$\begin{aligned} m_1 g \sin\alpha - \mu_k m_1 g \cos\alpha + T &= m_1 a, \\ m_2 g \sin\beta - \mu_k m_2 g \cos\beta - m_2 a &= T. \end{aligned}$$

Finally, we'll substitute the expression for T into the top equation to obtain

$$m_1 g \sin\alpha - \mu_k m_1 g \cos\alpha + (m_2 g \sin\beta - \mu_k m_2 g \cos\beta - m_2 a) = m_1 a,$$

which can be rewritten as

$$m_1 g \sin\alpha - \mu_k m_1 g \cos\alpha + m_2 g \sin\beta - \mu_k m_2 g \cos\beta = (m_1 + m_2) a.$$

Since we know the values of m_1, m_2, μ_k, α, and β, we can calculate all the quantities on the left-hand side and solve for a.

Other types of problems

Each of the previous examples asked you to find the acceleration, but sometimes a problem might give you the acceleration and ask you to solve for a different unknown. Regardless of what you must solve for, you should always start with a diagram and a sum-of-the-forces template. Once these equations are in front of you, you'll be able to reason through the problem more easily.

Experiment

You remove the spring from a retractable pen, and from the spring you suspend an object of known mass—say a 100[g] chocolate bar. With a ruler, you measure how much the spring stretches in the process. What is the spring constant k?

Discussion

In previous sections we discussed the *kinematics* problem of finding an object's position $x(t)$ given its acceleration function $a(t)$, and given the initial conditions x_i and v_i. In this section we studied the *dynamics* problem, which involves drawing force diagrams and calculating the net force acting on an object. Understanding these topics means you fully understand Newton's equation $F = ma$, which is perhaps the most important equation in this book.

We can summarize the entire procedure for predicting the position of an object $x(t)$ from first principles in the following equation:

$$\frac{1}{m}\underbrace{\left(\sum \vec{F} = \vec{F}_{\text{net}}\right)}_{\text{dynamics}} = \underbrace{\vec{a}(t) \quad \overset{\vec{v}_i+\int dt}{\longrightarrow} \quad \vec{v}(t) \quad \overset{\vec{r}_i+\int dt}{\longrightarrow} \quad \vec{r}(t)}_{\text{kinematics}}.$$

The left-hand side calculates the net force acting on an object, which is the *cause* of acceleration. The right-hand side indicates how we can calculate the position vector $\vec{r}(t)$ starting from the acceleration and the initial conditions. If you know the forces acting on any object (rocks, projectiles, cars, stars, planets, etc.) then you can predict the object's motion using this equation, which is pretty cool.

4.5 Momentum

A collision between two objects creates a sudden spike in the contact force between them, which can be difficult to measure and quantify. It is not possible to use Newton's law $F = ma$ to predict the accelerations that occur during collisions. To predict the motion of the objects after the collision, we need a *momentum* calculation. According to the law of conservation of momentum, the total amount of momentum before and after the collision is the same. Once we know the momenta of the objects before the collision, it becomes possible to calculate their momenta after the collision, and from this determine their subsequent motion.

To illustrate the importance of momentum, consider the following situation. Say you have a 1[g] piece of paper and a 1000[kg] car moving at the same speed of 100[km/h]. Which of the two objects would you rather be hit by? Momentum, denoted $\vec{p}$, is the precise physical concept that measures the *quantity* of motion. An object of mass m moving with velocity $\vec{v}$ has a momentum of $\vec{p} \equiv m\vec{v}$. Momentum plays a key role in collisions. Your gut feeling about the piece of paper and the car is correct. The car weighs $1000 \times 1000 = 10^6$ times more than the piece of paper, so the car has 10^6 times more momentum when moving at the same speed. Colliding with the car will "hurt" one-million times more than colliding with the piece of paper, even though both objects approach have the same velocity.

In this section, we'll learn how to use the law of conservation of momentum to predict the outcomes of collisions.

Concepts

- m: the *mass* of the moving object
- $\vec{v}$: the *velocity* of the moving object
- $\vec{p} = m\vec{v}$: the *momentum* of the moving object
- $\sum \vec{p}_{\textrm{in}}$: the sum of the momenta of particles before a collision
- $\sum \vec{p}_{\textrm{out}}$: the sum of the momenta of particles after a collision

Definition

The momentum of a moving object is equal to the velocity of the object multiplied by its mass:

$$\vec{p} = m\vec{v} \qquad \text{[kg m/s]}.$$

If an object's velocity is $\vec{v} = 20\hat{\imath} = (20, 0)$[m/s] and its mass is 100[kg], then its momentum is $\vec{p} = 2000\hat{\imath} = (2000, 0)$[kg m/s].

Momentum is a vector quantity, and we will often need to convert momentum from the length-and-direction form into the component form:

$$\vec{p} = \|\vec{p}\|\angle\theta = (\|\vec{p}\|\cos\theta, \|\vec{p}\|\sin\theta) = (p_x, p_y).$$

The component form makes it easy to add and subtract vectors: $\vec{p}_1 + \vec{p}_2 = (p_{1x} + p_{2x}, p_{1y} + p_{2y})$. To express the final answer, we will need to convert the component form back to the length-and-direction form:

$$\|\vec{p}\| = \sqrt{p_x^2 + p_y^2}, \qquad \theta = \tan^{-1}\left(\frac{p_y}{p_x}\right).$$

Conservation of momentum

Newton's first law states that in the absence of acceleration ($\vec{a} = 0$), an object maintains a constant velocity. This becomes kind of obvious if you apply the logic of calculus: $\vec{a}$ is the change in $\vec{v}$, so if $\vec{a} = 0$ then $\vec{v}$ must be constant.

In the absence of acceleration, objects conserve their velocity: $\vec{v}_{\text{in}} = \vec{v}_{\text{out}}$. When we multiply both sides of this equation by the object's mass, we obtain an equivalent statement saying that objects conserve their momentum:

$$\vec{p}_{\text{in}} = m\vec{v}_{\text{in}} \;=\; m\vec{v}_{\text{out}} = \vec{p}_{\text{out}}.$$

More generally, for situations involving multiple moving objects, the *sum* of the momenta of all the objects stays constant even if the objects interact. This reasoning is useful when analyzing collisions, since it allows us to equate the sum of the momenta before and after the collision:

$$\sum \vec{p}_{\text{in}} = \sum \vec{p}_{\text{out}}. \tag{4.2}$$

Any momentum that goes into a collision must also come out. This equation expresses the law of conservation of momentum.

The law of conservation of momentum is one of the furthest-reaching laws of physics you will learn by studying mechanics. We discussed the conservation of momentum in the simple context of two colliding particles, but the law applies widely, to multiple particles, fluids, fields, and even collisions involving atomic particles described by quantum mechanics. The quantity of motion (a.k.a. momentum) cannot be created or destroyed—it can only be exchanged between systems.

Examples

Example 1 It's a rainy day, and from your balcony you throw—horizontally, at a speed of 10[m/s]—a piece of rolled-up carton with a mass of 0.4[g]. Shortly after it leaves your hand, the piece collides with a rain drop that weighs 2[g] and is falling straight down at a speed of 30[m/s]. What will the resulting velocity be if the two objects stick together after the collision?

The conservation of momentum equation says,

$$\vec{p}_{\text{in},1} + \vec{p}_{\text{in},2} = \vec{p}_{\text{out}}.$$

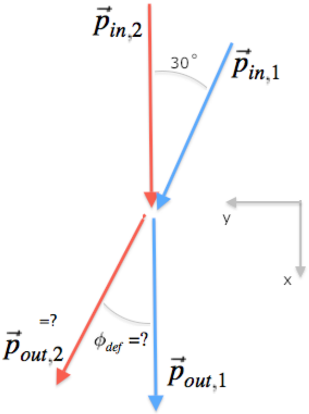
Plugging in the values, we obtain the equation

$$\begin{array}{ccccc} m_1\vec{v}_1 & + & m_2\vec{v}_2 & = & (m_1+m_2)\vec{v}_{\text{out}}, \\ 0.4\times(10,0) & + & 2\times(0,-30) & = & 2.4\times\vec{v}_{\text{out}}. \end{array}$$

Solving for $\vec{v}_{\text{out}}$ we find

$$\vec{v}_{\text{out}} = \frac{0.4(10,0)-2(0,30)}{2.4} = (1.666,-25.0) = 1.666\hat{\imath} - 25.0\hat{\jmath}.$$

Example 2: Hipsters on bikes Two hipsters on fixed-gear bikes are headed toward the same intersection. Both hipsters have a speed of 50[km/h]. The first hipster crosses the street at a diagonal of 30 degrees when the two bikers collide. Did anyone else see this coming? Apparently, the second hipster didn't, because the thick frames of his glasses were blocking his peripheral vision.



Let's look at the hipster moving in the straight line, and let's assume the combined weight of the hipster and his bike is 100[kg]. As for the street-crossing-at-30-degrees hipster, his weight combined with the weight of his bike frame (which is lighter and more expensive) totals 90[kg].

The story will continue in a moment, but first let's review the information I've given you so far:

$$\begin{aligned}\vec{p}_{\text{in},1} &= 90 \times 50\angle 30\\ &= 90(50\cos 30, 50\sin 30),\end{aligned}$$

$$\begin{aligned}\vec{p}_{\text{in},2} &= 100 \times 50\angle 0\\ &= (5000, 0),\end{aligned}$$

where the x-coordinate points down the street, and the y-coordinate is perpendicular to the street.

Surprisingly, nobody gets hurt in this collision. The bikers bump shoulder-to-shoulder and bounce off each other. The hipster who was trying to cross the street is redirected down the street, while the hipster travelling down the street is deflected to the side and rerouted onto a bike path. I know what you are thinking: couldn't they get hurt at least a little bit? Okay, let's say the whiplash from the shoulder-to-shoulder collision sends the hipsters' heads flying toward each other and smashes their glasses. There you have it.

Suppose the velocity of the first hipster after the collision is 60 [km/h]. What is the velocity and the deflected direction of the second hipster? As given above, the outgoing momentum of the first hipster is $\vec{p}_{\text{out},1} = (90 \times 60, 0)$, and we're looking to find $\vec{p}_{\text{out},2}$.

We can solve this problem with the conservation of momentum formula, which tells us that

$$\vec{p}_{\text{in},1} + \vec{p}_{\text{in},2} \;=\; \vec{p}_{\text{out},1} + \vec{p}_{\text{out},2}.$$

We know three of the above quantities, so we can solve for the remaining unknown vector by isolating it on one side of the equation:

$$\vec{p}_{\text{out},2} = \vec{p}_{\text{in},1} + \vec{p}_{\text{in},2} - \vec{p}_{\text{out},1},$$

$$\vec{p}_{\text{out},2} = 90(50\cos 30, 50\sin 30) + (5000, 0) - (90 \times 60, 0).$$

The x-component of the momentum $\vec{p}_{\text{out},2}$ is

$$p_{\text{out},2,x} = 90 \times 50\cos 30 + 5000 - 90 \times 60 = 3497.11,$$

and the y-component is $p_{\text{out},2,y} = 90 \times 50\sin 30 = 2250.$

The magnitude of the momentum of hipster 2 is given by

$$\|\vec{p}_{\text{out},2}\| = \sqrt{p^2_{\text{out},2,x} + p^2_{\text{out},2,y}} = 4158.39 \quad [\text{kg km/h}].$$

Note the unit of the momentum is not the standard choice [kg m/s]. That is fine. As long as you keep in mind which units you're using, it's not always necessary to convert to SI units.

The final velocity of hipster 2 is $v_{\text{out},2} = 4158.39/100 = 41.58$[km/h]. The deflection angle is obtained by

$$\phi_{\text{def}} = \tan^{-1}\!\left(\frac{p_{\text{out},2,y}}{p_{\text{out},2,x}}\right) = 32.76°.$$

Discussion

We previously defined the concept of momentum in terms of an object's velocity; but in fact, momentum can be traced to a concept more fundamental than velocity. If you go on to take more advanced physics classes, you'll learn about the *natural* variables—position and momentum $(\vec{x}, \vec{p})$—that describe the *state* of a particle. You'll also learn that the *real* form of Newton's second law is written in terms of momentum:

$$\vec{F} = \frac{d\vec{p}}{dt} \quad \text{for } m \text{ constant} \;\Rightarrow\quad \vec{F} = \frac{d(m\vec{v})}{dt} = m\frac{d\vec{v}}{dt} = m\vec{a}.$$

In most physics problems, objects will maintain a constant mass, so using $\vec{F} = m\vec{a}$ is perfectly fine.

The law of conservation of momentum follows from Newton's third law: for each force $\vec{F}_{12}$ exerted by Object 1 on Object 2, there exists a counter force $\vec{F}_{21}$ of equal magnitude and opposite direction, which is the force of Object 2 pushing back on Object 1. Earlier, I mentioned it is difficult to quantify the magnitude of the exact forces $\vec{F}_{12}$ and $\vec{F}_{21}$ that occur during a collision. Indeed, the amount of force suddenly shoots up as the two objects collide, then suddenly drops again. Complicated as these forces may be, we know that during the entire collision they obey Newton's third law. Assuming there are no other forces acting on the objects, we have

$$\vec{F}_{12} = -\vec{F}_{21} \quad \text{using the above} \Rightarrow \quad \frac{d\vec{p}_1}{dt} = -\frac{d\vec{p}_2}{dt}.$$

If we move the negative term to the left-hand side of the equation we obtain

$$\frac{d\vec{p}_1}{dt} + \frac{d\vec{p}_2}{dt} = 0 = \frac{d}{dt}\left(\vec{p}_1 + \vec{p}_2\right).$$

The second part of the equation implies that the quantity $(\vec{p}_1 + \vec{p}_2)$ is constant over time, and so $\vec{p}_{\text{in},1} + \vec{p}_{\text{in},2} = \vec{p}_{\text{out},1} + \vec{p}_{\text{out},2}$.

In this section, we saw how to use a momentum calculation to predict the motion of particles after a collision. In the next section we'll learn about *energy*, which is another useful concept for understanding and predicting the motion of objects.

Links

[Animations of simple collisions between objects]
http://en.wikipedia.org/wiki/Conservation_of_linear_momentum

4.6 Energy

Instead of thinking in terms of velocities $v(t)$ and motion trajectories $x(t)$, we can solve physics problems by using *energy* calculations. In this section, we'll precisely define different kinds of energies, and we'll learn the rules for converting one energy into another. The key idea to keep in mind is the principle of *total energy conservation*, which says that in any physical process, the sum of the initial energies is equal to the sum of the final energies.

Example

You drop a ball from a height h[m] and want to predict its speed just before it hits the ground. Through the kinematics approach, you would set up the general equation of motion,

$$v_f^2 = v_i^2 + 2a(y_f - y_i),$$

substitute $y_i = h$, $y_f = 0$, $v_i = 0$, and $a = -g$, and solve for the ball's final velocity at impact v_f. The answer is $v_f = \sqrt{2gh}$[m/s].

Alternately, we can use an energy calculation. The ball starts from a height h, which means it has $U_i = mgh$[J] of potential energy. As the ball falls, potential energy is converted into kinetic energy. Just before the ball hits the ground, its final kinetic energy is equal to the initial potential energy: $K_f = U_i$. Since the formula for kinetic energy is $K = \frac{1}{2}mv^2$[J], we have $\frac{1}{2}mv_f^2 = mgh$. We cancel the mass on both sides of the equation and solve for v_f to obtain $v_f = \sqrt{2gh}$[m/s].

Both methods of solving the example problem lead us to the same conclusion, but the energy reasoning is arguably more intuitive than blindly plugging values into a formula. In science, it is really important to know different ways of arriving at the same answer. Knowing about these alternate routes will allow you to check your answers and better understand concepts.

Concepts

Energy is measured in Joules [J] and it arises in several contexts:

- K = **kinetic energy**: the type of energy objects have by virtue of their motion
- W = **work**: the amount of energy an external force adds or subtracts from a system. Positive work corresponds to energy added to the system while negative work corresponds to energy withdrawn from the system.
- U_g = **gravitational potential energy**: the energy an object has by virtue of its position above the ground. We say this energy is *potential* because it is a form of *stored work*. Potential energy corresponds to the amount of work the force of gravity will add to an object when the object falls to the ground.
- U_s = **spring potential energy**: the energy stored in a spring when it is displaced (stretched or compressed) from its relaxed position

  There are many other kinds of energy—electrical energy, magnetic energy, sound energy, thermal energy, and so on. However, we'll limit our focus in this section to include only the *mechanical* energy concepts described above.

Formulas

Kinetic energy

An object of mass m moving at velocity $\vec{v}$ has a *kinetic energy* of

$$K = \frac{1}{2}m\|\vec{v}\|^2 \qquad \text{[J]}.$$

Note, the kinetic energy depends only on the speed $\|\vec{v}\|$ of the object and is not affected by the direction of motion.

Work

If an external force $\vec{F}$ acts on an object as it travels a distance $\vec{d}$, the *work* done by this force is

$$W = \vec{F} \cdot \vec{d} = \|\vec{F}\|\|\vec{d}\| \cos\theta \qquad \text{[J]}.$$

The second equality follows from the geometrical interpretation of the dot product $\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta$, where θ is the angle between $\vec{u}$ and $\vec{v}$.

If the force $\vec{F}$ acts in the same direction as the displacement $\vec{d}$, then the force will do positive work ($\cos(0°) = +1$) by adding energy to the system. If the force acts in the opposite direction to the direction of the displacement, the force will do negative work ($\cos(180°) = -1$) by withdrawing energy from the system.

Gravitational potential energy

An object raised to a height h above the ground has a gravitational potential energy given by

$$U_g(h) = mgh \qquad \text{[J]},$$

where m is the mass of the object and $g = 9.81$[m/s^2] is the gravitational acceleration on the surface of the Earth.

Spring potential energy

The potential energy stored in a spring when it is displaced by $\vec{x}$[m] from its relaxed position is given by

$$U_s = \frac{1}{2}k\|\vec{x}\|^2 \qquad \text{[J]},$$

where k[N/m] is the spring constant.

Note, it is irrelevant whether the spring is stretched or compressed: only the magnitude of the displacement matters $\|\vec{x}\|$.

Conservation of energy

Consider a system that starts from an initial state (i), undergoes some motion, and arrives at a final state (f). The law of conservation of energy states that **energy cannot be created or destroyed in any physical process**. The initial energy of the system plus the work that is *in*put into the system must equal the final energy of the system plus any work that is *out*put:

$$\sum E_i \quad + W_{\text{in}} \quad = \quad \sum E_f \quad + W_{\text{out}}.$$

The expression $\sum E_i$ corresponds to the sum of all the different types of energy the system contains in its initial state. Similarly, $\sum E_f$ corresponds to the sum of the final energies of the system. In mechanics, we consider three types of energy: kinetic energy, gravitational potential energy, and spring potential energy. Thus the conservation of energy equation in mechanics is

$$K_i + U_{gi} + U_{si} \quad + W_{\text{in}} \quad = \quad K_f + U_{gf} + U_{sf} \quad + W_{\text{out}}. \tag{4.3}$$

Usually, we're able to drop some of the terms in this lengthy expression. For example, we do not need to consider the spring potential energy U_s in physics problems that do not involve springs.

Explanations

Work and energy are measured in Joules [J]. Joules can be expressed in terms of other fundamental units:

$$[\text{J}] = [\text{N m}] = [\text{kg m}^2/\text{s}^2].$$

The first equality follows from the definition of work as force times displacement. The second equality comes from the definition of the Newton: $[\text{N}] = [\text{kg m/s}^2]$, which comes from $F = ma$.

Kinetic energy

A moving object has energy $K = \frac{1}{2}m\|\vec{v}\|^2$[J], called *kinetic* energy from the Greek word for motion, *kinema*.

Note that velocity $\vec{v}$ and speed $\|\vec{v}\|$ are not the same as energy. Suppose you have two objects of the same mass, and one is moving two times faster than the other. The faster object will have twice the velocity of the slower object, and four times more kinetic energy.

Work

When hiring movers to help you move, you must pay them for the *work* they do. Work is the product of the amount of force needed for the move and the distance of the move. When the move requires a lot of force, more work will be done. And the bigger the displacement (think moving to the South Shore versus moving next door), the more money the movers will ask for.

The amount of work done by a force $\vec{F}$ on an object that moves along some path p is given by

$$W = \int_p \vec{F}(x) \cdot d\vec{x}.$$

The integral account for the fact that the force's magnitude and direction might change along the path of motion.

If the force is constant and the displacement path is a straight line, the formula for work simplifies to

$$W = \int_0^d \vec{F} \cdot d\vec{x} = \vec{F} \cdot \int_0^d \, d\vec{x} = \vec{F} \cdot \vec{d} = \|\vec{F}\|\|\vec{d}\| \cos\theta.$$

Note the use of the dot product to obtain only the part of $\vec{F}$ that is pushing in the direction of the displacement $\vec{d}$. A force that acts perpendicular to the displacement produces no work, since it neither speeds nor slows the object's motion.

Potential energy is stored work

Some kinds of work are just a waste of time, like working at a job you despise. You work and you get your paycheque, but you don't learn anything useful at the end of the day. Other kinds of work leave you with some useful resource at the end of the work day—they grow your human potential.

In physics, we make a similar distinction. Some types of work, like work against friction, are called *dissipative* since they simply waste energy. Other kinds of work are called *conservative* since the work performed isn't lost but converted into *potential energy*.

The gravitational force and the spring force are conservative forces. Any work you do lifting an object into the air against the force of gravity is not lost but *stored* in the height of the object. By letting go of the object, you will get a full return on all the work performed while lifting the object. The energy will return in the form of kinetic energy since the object picks up speed during the fall.

The negative of the work done against a conservative force is called *potential energy*. For any conservative force $\vec{F}_?$, we can define the associated potential energy $U_?$ through the formula,

$$U_?(d) = -W_{\text{done}} = -\int_0^d \vec{F}_? \cdot d\vec{x}.$$

We'll discuss two specific examples of this general formula below: gravitational potential energy and spring potential energy. An object high in the air has a great potential to fall; similarly, compressing a spring by a certain distance gives it the potential to spring back to its normal position. Let's look at the exact formulas for these two cases.

Gravitational potential energy

The force of gravity is given by

$$\vec{F}_g = -mg\hat{\jmath}.$$

The direction of the gravitational force pulls downward, toward the centre of the Earth.

The gravitational potential energy of lifting an object from a height of $y = 0$ to a height of $y = h$ is given by

$$\begin{aligned} U_g(h) \equiv -W_{\text{done}} &= -\int_0^h \vec{F}_g \cdot d\vec{y} = -\int_0^h (-mg\hat{\jmath}) \cdot \hat{\jmath}\, dy \\ &= mg\int_0^h 1\, dy = mgy\Big|_{y=0}^{y=h} \\ &= mgh \qquad \text{[J]}. \end{aligned}$$

Spring energy

The force of a spring when stretched a distance $\vec{x}$[m] from its natural position is given by

$$\vec{F}_s(\vec{x}) = -k\vec{x}.$$

The potential energy stored in a spring as it is compressed from $y = 0$ to $y = x$[m] is given by

$$\begin{aligned} U_s(x) \equiv -W_{\text{done}} &= -\int_0^x \vec{F}_s(y) \cdot d\vec{y} \\ &= -\int_0^x (-ky)\, dy = k\int_0^x y\, dy = k\frac{1}{2}y^2\Big|_{y=0}^{y=x} \\ &= \frac{1}{2}kx^2 \qquad \text{[J]}. \end{aligned}$$

Conservation of energy

Energy cannot be created or destroyed. It can only be transformed from one form to another. If no external forces act on the system, then the system obeys the *conservation of energy* equation:

$$\sum E_i \;=\; \sum E_f.$$

If any external forces like friction do work on the system, we must account for their energy contributions:

$$\sum E_i \ + \ W_{\text{in}} = \sum E_f, \quad \text{or} \quad \sum E_i = \sum E_f \ + \ W_{\text{out}}.$$

The conservation of energy is one of the most important equations you will find in this book. It allows you to solve complicated problems by simply accounting for all the different kinds of energy involved in a system.

Examples

Banker dropped

An investment banker is dropped (from rest) from a building 100[m] tall. What is his speed when he hits the ground?

We start from

$$\begin{aligned} \sum E_i &= \sum E_f, \\ K_i + U_i &= K_f + U_f, \end{aligned}$$

and plug in the numbers to get

$$0 + m \times 9.81 \times 100 = \frac{1}{2} m v^2 + 0.$$

We cancel the mass m from both sides of the equation and are left with

$$9.81 \times 100 = \frac{1}{2} v_f^2.$$

Solving for v_f in the above equation, we find the banker will be falling at $v_f = \sqrt{2 \times 9.81 \times 100} = 44.2945$[m/s] when he hits the ground. This is about 160[km/h]. Ouch! That will definitely hurt.

Bullet speedometer

An incoming bullet at speed v hits a block of mass M which is suspended in the air. Use conservation of momentum and conservation of energy principles to find the speed v of the bullet if the block rises to a height h after being hit by the bullet.

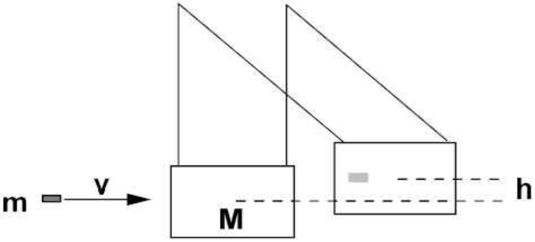


Figure 4.9: A bullet of mass m hits a block of mass M and causes the mass to swing up and to the right. Find the incoming velocity of the bullet v if the block-plus-bullet system rises by a height h after the impact.

First we apply the conservation of momentum principle to find the block's horizontal speed and its mass just after the bullet hits:

$$\begin{aligned} \vec{p}_{\text{in},m} + \vec{p}_{\text{in},M} &= \vec{p}_{\text{out}}, \\ mv + 0 &= (m+M)v_{\text{out}}. \end{aligned}$$

Thus, the block's velocity just after the bullet's impact is $v_{\text{out}} = \frac{mv}{M+m}$.

Next, we go to the conservation of energy principle to relate the initial kinetic energy of the block-plus-bullet to the height h by which the block rises:

$$\begin{aligned} K_i + U_i &= K_f + U_f, \\ \frac{1}{2}(M+m)v_{\text{out}}^2 + 0 &= 0 + (m+M)gh. \end{aligned}$$

Isolate v_{out} in the above equation and set it equal to the v_{out} we obtained from the momentum calculation:

$$v_{\text{out}} = \frac{mv}{M+m} = \sqrt{2gh} = v_{\text{out}}.$$

We can use this equation to find the speed of the incoming bullet:

$$v = \frac{M+m}{m}\sqrt{2gh}.$$

Incline and spring

A block of mass m is released from rest at point (A), located on top of an incline at coordinate $y = y_i$. The block slides down the frictionless incline to the point (B) $y = 0$. The coordinate $y = 0$ corresponds to the relaxed length of a spring with spring constant k. The block then compresses the spring all the way to point (C), corresponding to $y = y_f$, where the block comes to rest again. The angle of the incline's slope is θ.

What is the speed of the block at $y = 0$? Find the value of y_f, the compression of the spring when the block stops. Bonus points if you can express your answer for y_f in terms of Δh, the difference in height between points (A) and (C).

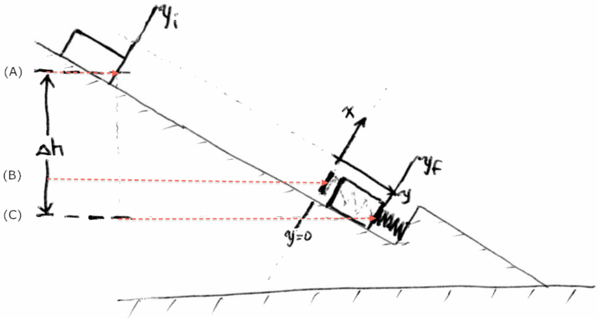


Figure 4.10: A block is released from the point (A) and slides down a frictionless incline to the point (B). The motion of the block is then slowed by a spring at the bottom of the incline. The block comes to rest at the point (C), after the spring is compressed by a length y_f.

Essentially, we have two problems: the block's motion from (A) to (B) in which its gravitational potential energy is converted into kinetic energy; and the block's motion from (B) to (C), in which all its energy is converted into spring potential energy.

There is no friction in either movement, so we can use the conservation of energy formula:

$$\sum E_i \quad = \quad \sum E_f.$$

For the block's motion from (A) to (B), we have

$$K_i + U_i = K_f + U_f.$$

The block starts from rest, so $K_i = 0$. The difference in potential energy is equal to mgh, and in this case the block is $|y_i|\sin\theta$ [m] higher at (A) than it is at (B), so we write

$$0 + mg|y_i|\sin\theta = \frac{1}{2}mv_B^2 + 0.$$

In the formula above, we assume the block has zero gravitational potential energy at point (B). The potential energy at point (A) is $U_i = mgh = mg|y_i - 0|\sin\theta$ *relative* to point (B), since point (A) is $h = |y_i - 0|\sin\theta$ metres higher than point (B).

Solving for v_B in this equation answers the first part of our question:

$$v_B = \sqrt{2g|y_i|\sin\theta}.$$

Now for the second part of the block's motion. The law of conservation of energy dictates that

$$K_i + U_{gi} + U_{si} = K_f + U_{gf} + U_{sf},$$

where i now refers to the moment (B), and f refers to the moment (C). Initially the spring is uncompressed, so $U_{si} = 0$. By the end of the motion, the spring is compressed by a total of $\Delta y = |y_f - 0|$[m], so its spring potential energy is $U_{sf} = \frac{1}{2}k|y_f|^2$. We choose the height of (C) as the reference potential energy; thus $U_{gf} = 0$. Since the difference in gravitational potential energy is $U_{gi} - U_{gf} = mgh = |y_f - 0|\sin\theta$, we can complete the entire energy equation:

$$\frac{1}{2}mv_B^2 + mg|y_f|\sin\theta + 0 = 0 + 0 + \frac{1}{2}k|y_f|^2.$$

Assuming the values of k and m are given, and knowing v_B from the first part of the question, we can solve for $|y_f|$ in the above equation.

To obtain the answer $|y_f|$ in terms of Δh, we'll use $\sum E_i = \sum E_f$ again, but this time i will refer to moment (A) and f to moment (C). The conservation of energy equation tells us $mg\Delta h = \frac{1}{2}k|y_f|$, from which we obtain $|y_f| = \frac{2mg\Delta h}{k}$.

Energy lost to friction

You place a block of mass 50[kg] on an incline. The force of friction between the block and the incline is 30[N]. The block slides for 200[m] down the incline. The incline's slope is $\theta = 30°$ making the block's total vertical displacement $200 \sin 30 = 100$[m]. What is the block's speed as it reaches the bottom of the incline?

This is a problem in which initial energies are converted into a combination of final energies and *lost* work:

$$\sum E_i = \sum E_f + W_{\text{lost}}.$$

The term W_{lost} represents energy lost due to friction.

A better way of describing this situation is that **a negative amount of work is done on the block**:

$$\sum E_i + \underbrace{W_{\text{done}}}_{\text{negative}} = \sum E_f.$$

The quantity W_{done} is negative because the friction force acts on the object in the opposite direction of the object's motion:

$$W_{\text{done}} = \vec{F} \cdot \vec{d} = \|\vec{F}_f\|\|\vec{d}\| \cos(180°) = -F_f\|\vec{d}\|,$$

where $\vec{d}$ is the sliding distance of 200[m] over which the friction acts.

We substitute the value of W_{done} into the conservation of energy equation:

$$K_i + U_i + W_{\text{done}} = K_f + U_f,$$

$$0 + mgh + (-F_f|d|) = \frac{1}{2}mv_f^2 + 0.$$

Note we used the formula $mgh = U_i - U_f$ for the difference in gravitational potential energy.

Since we're told $F_f = 30$[N], we can calculate $W_{\text{done}} = W_{\text{friction}} = -30$[N] $\times$ 200[m] $= -6000$[J]. Substituting all known values, we find

$$0 + 50 \times 9.81 \times 100 - 6000 = \frac{1}{2}(50)v_f^2 + 0,$$

which we can solve for v_f.

Discussion

It's useful to describe physical situations in terms of the energies involved. The law of conservation of energy allows us to use simple "energy accounting" principles to calculate the values of unknown quantities.

4.7 Uniform circular motion

This section covers the circular motion of objects. Circular motion differs from linear motion, and we'll need to learn new techniques and concepts specifically used to describe circular motion.

Imagine a rock of mass m attached to the end of a rope and swinging around in a horizontal circle. The rock flies through the air at a constant speed of v_t[m/s], along a circular path of radius R[m], at a height h[m] above the ground. What is the tension T in the rope?

Consider a coordinate system with its x and y axes placed on the ground. At the centre of the circle of motion is the z-axis, which measures the height above the ground. In this (x, y, z) coordinate system, the rock's trajectory is described by the equation

$$\vec{r}(t) = (x(t), y(t), z(t)) = \left(R\cos\!\left(\frac{v_t}{R}\,t\right),\ R\sin\!\left(\frac{v_t}{R}\,t\right),\ h\right).$$

Do you agree with me that this expression looks somewhat complicated? Its complexity stems from the fact that the (x, y, z) coordinate system is not well-adapted for describing circular paths.

A new coordinate system

Instead of the usual coordinate system $\hat{x}, \hat{y}, \hat{z}$, which is static, we can choose a new coordinate system $\hat{t}, \hat{r}, \hat{z}$ that is "attached" to the rotating object. The new coordinate system consists of:

- $\hat{t}$: the *tangential* direction is the object's instantaneous direction of motion. The name comes from the Greek word for "touch" (imagine a straight line "touching" the circle).
- $\hat{r}$: the *radial* direction always points toward the centre of the circle of rotation.
- $\hat{z}$: the usual $\hat{z}$-direction

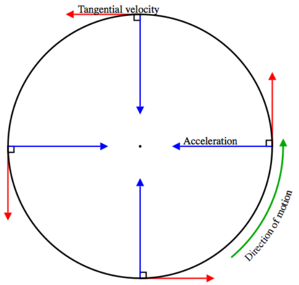
From a static observer's point of view, the tangential and radial directions constantly change their orientation as the object rotates around in a circle. From the rotating object's point of view, the tangential and radial directions are fixed. The tangential direction is always "forward" and the radial direction is always "to the side."

We can use the new coordinate system to describe the position, velocity, and acceleration of an object undergoing circular motion:

- $\vec{v} = (v_r, v_t)_{\hat{r}\hat{t}}$: the object's *velocity* expressed with respect to the $\hat{r}\hat{t}$-coordinate system
- $\vec{a} = (a_r, a_t)_{\hat{r}\hat{t}}$: the object's *acceleration*

The most important parameters of motion are the tangential velocity v_t, the radial acceleration a_r, and the radius of the circle of motion R. We have $v_r = 0$ since the motion is entirely in the $\hat{t}$-direction, and $a_t = 0$ because in this case we assume the tangential velocity v_t remains constant (*uniform* circular motion).

Radial acceleration



The defining feature of circular motion is the presence of an acceleration that acts perpendicularly to direction of motion. At each instant, the object wants to continue moving along the tangential direction, but the radial acceleration causes the object's velocity to change direction. This constant inward acceleration causes the object to follow a circular path.

The radial acceleration a_r of an object moving in a circle of radius R with a tangential velocity v_t is given by

$$a_r = \frac{v_t^2}{R}\,.$$

This important equation relates the three key parameters of circular motion.

According to Newton's second law $\vec{F} = m\vec{a}$, an object's radial acceleration must be caused by a *radial force*. We can calculate the magnitude of this radial force F_r as follows:

$$F_r = ma_r = m\frac{v_t^2}{R}\,.$$

This formula connects the observable aspects of a circular motion v_t and R with the motion's cause: the force F_r, which always acts toward the centre of rotation.

To phrase it another way, we can say circular motion *requires* a radial force. From now on, when you see an object in circular motion, you can try to visualize the radial force that is causing the circular motion.

In the rock-on-a-rope example introduced in the beginning of the section (page 219), circular motion is caused by the tension of a rope that always acts in the radial direction (toward the centre of rotation). We're now in a position to calculate the value of the tension T in the rope using the equation

$$F_r = T = ma_r \qquad \Rightarrow \qquad T = m\frac{v_t^2}{R}\,.$$

The heavier the object and the faster it goes, the higher the tension in the rope. Inversely, the bigger the circle's radius, the less tension is required for the same v_t.

Example

During a student protest, a young activist named David is stationed on the rooftop of a building of height 12[m]. A mob of blood-thirsty neoconservatives is slowly approaching his position, determined to lynch him because of his leftist views. David has assembled a makeshift weapon by attaching a 0.3[kg] rock to the end of a shoelace of length 1.5[m]. The maximum tension the shoelace can support is 500[N]. What is the maximum tangential velocity $\max\{v_t\}$ the shoelace can support? What is the projectile's maximum *range* when it is launched from the roof?

The first part of the question is answered with the $T = m\frac{v_t^2}{R}$ formula: $\max\{v_t\} = \sqrt{\frac{RT}{m}} = \sqrt{\frac{1.5\times 500}{0.3}} = 50$[m/s]. To answer the second question, we must solve for the distance travelled by a projectile with initial velocity $\vec{v}_i = (v_{ix}, v_{iy}) = (50, 0)$[m/s], launched from $\vec{r}_i = (x_i, y_i) = (0, 12)$[m]. First, solve for the total time of flight $t_f = \sqrt{2 \times 12/9.81} = 1.56$[s]. Then find the range of the rock by multiplying the projectile's horizontal speed by the time of flight $x(t_f) = 0 + v_{ix}t_f = 50 \times 1.56 = 78.20$[m].

After carrying out these calculations on a piece of paper, David starts to spin-up the rock and waits for the neocons to come into range.

Circular motion parameters

It's time to introduce some further terminology that will help us describe circular motion:

- $C = 2\pi R$[m]: the *circumference* of the circle of motion
- T: the *period* of the motion. The time T is how long it takes for the object to complete one full circle. This period is measured in seconds [s].

- $f = \frac{1}{T}$: the frequency of rotation. The frequency describes the number of turns the object completes in one second. Frequency is measured in Hertz [Hz]=[1/s]. We sometimes describe the frequency of rotation in *revolutions per minute* (RPM).
- $\omega \equiv \frac{v_t}{R} = 2\pi f$: the *angular velocity* describes how fast the object is rotating. Angular velocity is measured in [rad/s].

Recall that a circle of radius R has a circumference $C = 2\pi R$. The period T is defined as how long it takes the object to complete one full turn around the circle:

$$T = \frac{\text{distance}}{\text{speed}} = \frac{C}{v_t} = \frac{2\pi R}{v_t},$$

where $C = 2\pi R$ is the total distance travelled to compete one turn, and v_t is the velocity of the object along the curve. The object completes one full turn every T seconds.

There is another way to describe circular motion by referencing an object's *frequency* of rotation:

$$f = \frac{1}{T} = [\text{Hz}].$$

The frequency of an object in circular motion tells you how many turns the object completes in one second. If the object completes one turn in $T = 0.2$[s], then the motion's frequency is $f = \frac{1}{0.2} = 5$[Hz], or $f = 60 \times 5 = 300$[RPM].

The most natural parameter for describing rotation is in terms of *angular velocity* ω[rad/s]. We know one full turn corresponds to an angle of rotation of 2π[rad], so angular velocity is obtained by dividing 2π by the time it takes to complete one turn:

$$\omega = \frac{2\pi}{T} = 2\pi f = \frac{v_t}{R}.$$

The angular velocity ω is useful because it describes the speed of a circular motion without any reference to the radius. If we know the angular velocity of an object is ω, we can obtain the tangential velocity by multiplying angular velocity times the radius: $v_t = R\omega$[m/s].

You'll be asked to compute some angular velocities in the upcoming examples.

Bicycle odometer

Imagine you place a small speed detector gadget on one of the spokes of your bicycle's front wheel. Your bike's wheels have a radius $R = 14$[in], and the gadget is attached at a distance $\frac{3}{4}R$[m] from the wheel's centre. Find the wheel's angular velocity ω, period T, and frequency f of rotation when the bicycle's speed relative to the ground is 40[km/h]. What is the tangential velocity v_t of the detector gadget?

The bicycle's velocity relative to the ground $v_{\text{bike}} = 40$[km/h] is equal to the tangential velocity of the rim of the wheel:

$$v_{\text{bike}} = v_{\text{rim}} = 40[\cancel{\text{km}}/\cancel{\text{h}}] \times \frac{1000[\text{m}]}{1[\cancel{\text{km}}]} \times \frac{1[\cancel{\text{h}}]}{3600[\text{s}]} = 11.11[\text{m/s}].$$

We can find the wheel's angular velocity using $\omega = \frac{v_{\text{rim}}}{R}$ and the radius of the wheel $R = 14[\text{in}] = 0.355[\text{m}]$. We obtain $\omega = \frac{11.11}{0.355} = 31.24$[rad/s]. From here we calculate $T = \frac{2\pi}{\omega} = 0.20$[s] and $f = \frac{1}{0.20} = 5$[Hz]. Finally, to compute the gadget's tangential velocity, multiply the wheel's angular velocity ω by its radius of rotation: $v_{\text{det}} = \omega \times \frac{3}{4}R = 8.333$[m/s].

Rotation of the Earth

It takes exactly 23 hours, 56 minutes and 4.09 seconds for the Earth to compete one full turn (2π radians) around its axis of rotation. What is the Earth's angular velocity? What is the tangential speed of a person standing in Montreal, at a latitude of $45°$?

We can find ω by carrying out a simple conversion:

$$\frac{2\pi[\text{rad}]}{1[\cancel{\text{day}}]} \cdot \frac{1[\cancel{\text{day}}]}{23.93447[\cancel{\text{h}}]} \cdot \frac{1[\cancel{\text{h}}]}{3600[\text{s}]} = 7.2921 \times 10^{-5}[\text{rad/s}].$$

The radius of the trajectory traced by someone located at a latitude of $45°$ is given by $r = R\cos(45°) = 4.5025 \times 10^{6}$[m], where $R = 6.3675 \times 10^{6}$[m] is the radius of the Earth. Though it may not feel like you're moving, you are actually hurtling through space at a speed of

$$v_t = r\omega = 4.5025 \times 10^6 \times 7.2921 \times 10^{-5} = 464.32[\text{m/s}],$$

which is equal to 1671.56[km/h]. Imagine that! You can attempt to present this fact if you are ever stopped by the cops for a speeding infraction: "Yes officer, I was doing 130[km/h], but this is really a negligible speed relative to the 1671[km/h] the Earth is doing around its axis of rotation."

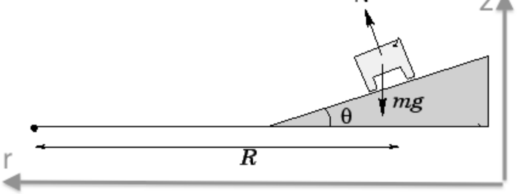
Three dimensions

For some problems involving circular motion, we'll need to consider the z-direction in the force diagram. In these cases, the best approach is to draw the force diagram as a cross section that is perpendicular to the tangential direction. Your diagram should show the $\hat{r}$ and $\hat{z}$ axes.

Using the force diagram, you can find all forces in the radial and vertical directions, as well as solve for accelerations a_r, a_z. Remember, you can always use the relation $a_r = \frac{v_t^2}{R}$, which connects the value of a_r with the tangential velocity v_t and the radius of rotation R.

Example Japanese people of the future design a giant racetrack for retired superconducting speed trains. The shape of the race track is a big circle with radius $R = 3$[km]. Because the trains are magnetically levitated, there is no friction between the track and the train $\mu_s = 0, \mu_k = 0$. What is the bank angle required for the racetrack so trains moving at a speed of exactly 400[km/h] will stay on the track without moving laterally?

We begin by drawing a force diagram which shows a cross section of the train in the $\hat{r}$ and $\hat{z}$ directions. The bank angle of the racetrack is θ. This is the unknown we're looking for. Because of the frictionless-



ness of levitated superconducting suspension, there cannot be any force of friction F_f. Therefore, the only forces acting on the train are its weight $\vec{W}$ and the normal force $\vec{N}$.

The next step is to write two force equations that represent the $\hat{r}$ and $\hat{z}$ directions:

$$\begin{aligned} \sum F_r = N\sin\theta = ma_r = m\frac{v_t^2}{R} \quad &\Rightarrow \quad N\sin\theta = m\frac{v_t^2}{R}, \\ \sum F_z = N\cos\theta - mg = 0 \qquad\quad &\Rightarrow \quad N\cos\theta = mg. \end{aligned}$$

Note how the normal force $\vec{N}$ is split into two parts: the vertical component counterbalances the train's weight so it doesn't slide down the track. The component of $\vec{N}$ in the $\hat{r}$-direction is the force that causes the train's rotational motion.

We want to solve for θ in the above equations. It's a common trick to solve equations containing multiple trigonometric functions by dividing one equation by the other. Doing this, we obtain

$$\frac{N\sin\theta}{N\cos\theta} = \frac{m\frac{v_t^2}{R}}{mg} \quad \Rightarrow \quad \tan\theta = \frac{v_t^2}{Rg}\,.$$

The final answer is $\theta = \tan^{-1}\!\left(\frac{v_t^2}{gR}\right) = \tan^{-1}\!\left(\frac{(400\times\frac{1000}{3600})^2}{9.81\times 3000}\right) = 22.76°$. If the angle were any steeper, the trains would fall toward the track's centre. If the bank angle were any shallower, the trains would fly off to the side. The angle $22.76°$ is just right.

Discussion

Radial acceleration

In the kinematics section we studied problems involving *linear acceleration*, in which an acceleration a acted in the direction of the velocity, causing a change in the magnitude of the velocity v.

Circular motion deals with a different situation in which the object's speed $\|\vec{v}\|$ remains constant while its velocity $\vec{v}$ changes direction. At each point along the circle, the object's velocity points along the tangential direction; simultaneously, the radial acceleration pulls the object inwards, causing it to rotate.

Another term for radial acceleration is *centripetal* acceleration, which literally means "tending toward the centre."

Nonexistence of the centrifugal force

When a car makes a left turn, the passenger riding shotgun will feel pushed toward the right, into the passenger door. It would be erroneous to attribute this effect to *centrifugal force*, which acts away from the centre of rotation. In fact, no force is directly responsible for the feeling of being flung out of a car during a sharp turn.

The passenger is pushed into the door because of Newton's first law, which says that in the absence of external forces, an object will continue moving in a straight line. Since the initial motion occurs in the forward direction, the passenger's body will naturally want to continue moving in that direction. The force of the car door on the passenger is necessary to keep her in the circular trajectory. If it weren't for the force of the door, she'd be flying straight out!

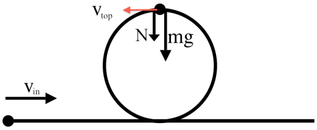
Radial forces do no work

An interesting property of radial forces is that they perform zero work. Recall that the work done by a force $\vec{F}$ during a displacement $\vec{d}$ is computed using the dot product $W = \vec{F} \cdot \vec{d}$. For circular motion, displacement is always in the $\hat{t}$-direction, while radial force acts in the $\hat{r}$-direction, making the dot product of the two vectors zero. Thus, the effects of radial forces do not increase the object's speed—they only act to change the direction of the velocity.

Exercise

Staying in touch

A racetrack features a vertical loop of radius 6.6[m]. A motorcyclist is about to attempt a loop-de-loop. What is the minimum speed v_{in} the motorcyclist needs in order to enter the ramp and drive all the way around the vertical loop? Bear in mind, the motorcyclist will lose contact with the track at the top of the ramp if the magnitude of the normal force drops to zero.



A radial force equal to $F_r = m\frac{v_{\text{top}}^2}{6.6}$ is required to keep the motorcycle in the loop. Assuming $N = 0$ at the top of the loop, the radial force at the top is given by $F_r = mg + N = mg + 0 = m\frac{v_{\text{top}}^2}{6.6}$, which we solve to find $v_{\text{top}} = \sqrt{6.6g}$. Next, we use conservation of energy $\frac{1}{2}mv_{\text{in}}^2 = m \times g \times 2 \times 6.6 + \frac{1}{2}mv_{\text{top}}^2$ to find $v_{\text{in}} = \sqrt{4 \times g \times 6.6 + 6.6 \times g} = \sqrt{5 \times 6.6 \times g} = 18$[m/s].

Links

[Loop-de-loop with a car]
http://www.youtube.com/watch?v=wiZoVAZGgsw

4.8 Angular motion

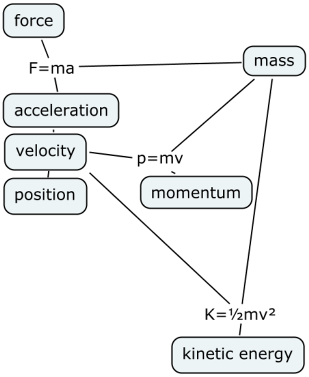
We will now study the physics of objects in rotation. Rotational motion is exemplified by spinning disks, rotating bicycle wheels, spinning footballs, and spinning figure skaters, among other spinning things.

As you'll see shortly, the basic concepts we'll use to describe angular motion are directly analogous to the concepts of linear motion: position, velocity, acceleration, force, momentum, and energy.

Review of linear motion

It will be helpful to begin with a quick review of the concepts and formulas used to describe the linear motion of objects.

The linear motion of an object is described by its position $x(t)$, velocity $v(t)$, and acceleration $a(t)$ as functions of time. The position function tells you where the object is, the velocity tells you how fast it is moving, and the acceleration measures the change in the object's velocity.



The motion of objects is governed by Newton's first and second laws. In the absence of external forces, objects will maintain a uniform velocity (UVM), which corresponds to the equations of motion $x(t) = x_i + v_i t$ and $v(t) = v_i$. If a net force $\vec{F}$ acts on the object, the force will cause the object to accelerate. We obtain the magnitude of this acceleration with the formula $F = ma$. A constant force acting on an object will produce a constant acceleration (UAM), which corresponds to the equations of motion $x(t) = x_i + v_i t + \frac{1}{2}at^2$ and $v(t) = v_i + at$.

We also learned how to quantify the *momentum* $\vec{p} = m\vec{v}$ and the *kinetic energy* $K = \frac{1}{2}mv^2$ of moving objects. The momentum vector is the natural measure of the "quantity of motion," which plays a key role in collisions. The kinetic energy measures how much energy the object has by virtue of its motion.

An object's mass m is also an important factor in many physics equations. In the equation $F = ma$, the mass m measures the object's *inertia*—the object's resistance to being moved. The object's mass also appears in the formulas for momentum and kinetic energy; the heavier the object, the larger its momentum and its kinetic energy will be.

Concepts

We're ready to introduce the new concepts for describing the angular motion of objects.

- The kinematics of rotating objects is described in terms of angular quantities:

  ▷ $\theta(t)$[rad]: the object's angular position
  ▷ $\omega(t)$[rad/s]: the object's angular velocity
  ▷ $\alpha(t)$[rad/s^2]: the object's angular acceleration

- I[kg m^2]: the *moment of inertia* tells you how difficult it is to make the object turn. The quantity I plays the same role in angular motion as the mass m plays in linear motion.
- $\mathcal{T}$[N m]: *torque* measures angular force. Torque is the cause of angular acceleration. The angular equivalent of Newton's second law $\sum F = ma$ is given by the equation $\sum \mathcal{T} = I\alpha$. This law states that applying an angular force (torque) $\mathcal{T}$ will produce an amount of angular acceleration α that is inversely proportional to the object's moment of inertia I.
- $L = I\omega$[kg m^2/s]: the *angular momentum* of a rotating object describes the "quantity of rotational motion."
- $K_r = \frac{1}{2}I\omega^2$[J]: the *angular* or *rotational* kinetic energy quantifies the amount of energy an object has by virtue of its rotational motion.

Formulas

Angular kinematics

Instead of talking about position x, velocity v, and acceleration a, for angular motion we will use the angular position θ, angular velocity ω, and angular acceleration α. Except for this change of ingredients, the "recipe" for finding the equations of motion remains the same:

$$\alpha(t) \quad \overset{\omega_i + \int dt}{\longrightarrow} \quad \omega(t) \quad \overset{\theta_i + \int dt}{\longrightarrow} \quad \theta(t).$$

Given the knowledge of an object's angular acceleration $\alpha(t)$, its initial angular velocity ω_i, and its initial angular position θ_i, we can use integration to find the equation of motion $\theta(t)$ that describes the angular position of the rotating object at all times.

Though this recipe can be applied to any form of angular acceleration function, you are only *required* to know the equations of motion for two special cases: the case of constant angular acceleration $\alpha(t) = \alpha$, and the case of zero angular acceleration $\alpha(t) = 0$. These are the angular analogues of *uniform acceleration motion* and *uniform velocity motion* we studied in the kinematics section.

The equations that describe *uniformly accelerated angular motion* are

$$\begin{aligned}\alpha(t) &= \alpha,\\ \omega(t) &= \alpha t + \omega_i,\\ \theta(t) &= \frac{1}{2}\alpha t^2 + \omega_i t + \theta_i,\\ \omega_f^2 &= \omega_i^2 + 2\alpha(\theta_f - \theta_i).\end{aligned}$$

Note how the form of the equations is *identical* to the linear UAM equations. This should come as no surprise since both sets of equations are obtained using the same integration procedure.

The equations of motion for *uniform velocity angular motion* are

$$\begin{aligned}\alpha(t) &= 0,\\ \omega(t) &= \omega_i,\\ \theta(t) &= \omega_i t + \theta_i.\end{aligned}$$

Relation to linear quantities

The angular quantities θ, ω, and α are the natural parameters for describing the motion of rotating objects. In certain situations, however, we may want to relate these angular quantities to linear quantities like distance, velocity, and linear acceleration. The connection between angular and linear quantities can be established by multiplying the angular quantity by the radius of motion:

$$d = R\theta, \qquad v = R\omega, \qquad a = R\alpha.$$

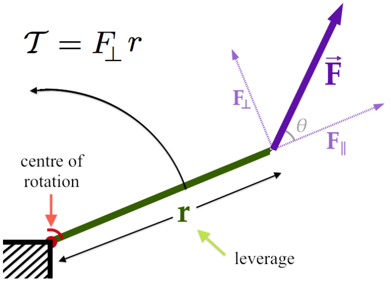
For example, suppose you have a spool of network cable with radius 20[cm], and you need to measure a length of 20[m] to connect your computer to your neighbour's computer. How many turns of the spool are needed to unwind 20[m] of cable? To find out, we can solve for θ in the formula $d = R\theta$ and obtain $\theta = 20/0.2 = 100$[rad], which corresponds to 15.9 turns.

Torque

Torque is angular force. In order to make an object rotate, you must exert a torque on it. Torque is measured in Newton metres [N m].
The torque produced by a force $\vec{F}$ is given by

$$\mathcal{T} = F_{\perp}\, r = \|\vec{F}\| \sin\theta\, r,$$

where r is the distance from the centre of rotation where the force is applied. Note that only the $F_{\perp}$ component of the force creates a torque. You can think of the distance r as *leverage*; even a small amount of force can produce a lot of torque if it acts far away from the centre.



To understand the meaning of the torque equation, you should stop reading for a moment and go experiment with a door. When you push on the door near the hinges, it takes a lot more force to make it move than when you push the edge of the door farthest from the hinges. The more leverage r you have, the more torque you'll produce. Also, if you pull the door's handle away from its hinges (as if trying to pull the door out of the wall), your force will have an $F_{||}$ component, but no $F_{\perp}$ component, so no matter how hard you pull, you will not cause the door to move.

As a standard convention, we use positive numbers to describe torques that produce counter-clockwise rotation, and negative numbers to describe torques that cause clockwise rotation.

The relationship between torque and force can also be used in the other direction. If a motor produces a torque of $\mathcal{T}$[N m] and is attached to a chain wheel of radius R, then the tension in the chain will be

$$T = F_{\perp} = \mathcal{T}/R \qquad [\text{N}].$$

With this equation you can compute the maximum pulling force produced by a car. You'll need to look up the value of the maximum torque produced by the car at the drive wheels, then divide by the radius of the wheels.

Moment of inertia

An object's momentum of inertia describes how difficult it is to cause the object to rotate:

$$I = \{ \text{ how difficult it is to make an object turn } \}.$$

The calculation describing the moment of inertia accounts for the mass distribution of the object. An object with most of its mass close to its centre will have a smaller moment of inertia, whereas objects with masses far from their centres will have larger moments of inertia.

The formula for calculating the moment of inertia is

$$I = \sum m_i r_i^2 = \int_{\text{obj}} r^2 \, dm \qquad [\text{kg m}^2].$$

The contribution of each piece of the object's mass dm to the total moment of inertia is proportional to the squared distance of that piece from the object's centre, hence the units [kg m^2].

We rarely use the integral formula to calculate objects' moments of inertia. Most physics problems you'll be asked to solve will involve geometrical shapes for which the moment of inertia is given by simple formulas:

$$I_{\text{disk}} = \frac{1}{2}mR^2, \; I_{\text{ring}} = mR^2, \; I_{\text{sphere}} = \frac{2}{5}mR^2, \; I_{\text{sph.shell}} = \frac{2}{3}mR^2.$$

When you learn more about calculus (Chapter 5), you will be able to derive each of the above formulas on your own. For now, just try to remember the formulas for the inertia of the disk and the ring, as they are likely to show up in problems.

The quantity I plays the same role in the equations of angular motion as the mass m plays in the equations of linear motion.

Torques cause angular acceleration

Recall Newton's second law $F = ma$, which describes the amount of acceleration produced by a given force acting on an object. The angular analogue of Newton's second law is expressed as

$$\mathcal{T} = I\alpha.$$

This equation indicates that the angular acceleration produced by the torque $\mathcal{T}$ is inversely proportional to the object's moment of inertia. Torque is the cause of angular acceleration.

Angular momentum

The angular momentum of a spinning object measures the "amount of rotational motion." The formula for the angular momentum of an object with moment of inertia I rotating at an angular velocity ω is

$$L = I\omega \qquad [\text{kg m}^2/\text{s}].$$

In the absence of external torques, an object's angular momentum is a conserved quantity:

$$L_{\text{in}} = L_{\text{out}}.$$

This property is similar to the way momentum $\vec{p}$ is a conserved quantity in the absence of external forces.

Rotational kinetic energy

The kinetic energy of a rotating object is calculated as

$$K_r = \frac{1}{2}I\omega^2 \qquad [\text{J}].$$

This expression is the rotational analogue to the linear kinetic energy formula $K = \frac{1}{2}mv^2$.

The amount of work produced by a torque $\mathcal{T}$ applied during an angular displacement of θ is given by

$$W = \mathcal{T}\theta \qquad [\text{J}].$$

With the equations above, we can now include the energy and work of rotating objects in our conservation-of-energy calculations.

Examples

Rotational UVM

A disk is spinning at a constant angular velocity of 12[rad/s]. How many turns will the disk complete in one minute?

Since the angular velocity is constant, we can use the equation $\theta(t) = \omega t + \theta_i$ to find the disk's total angular displacement after one minute. We obtain $\theta(60) = 12 \times 60 = 720$[rad]. To find the number of turns, divide this number by 2π and obtain 114.6[turns].

Rotational UAM

A solid disk of mass 20[kg] and radius 30[cm] is initially spinning with an angular velocity of 20[rad/s]. A brake pad applied to the edge of the disk produces a friction force of 60[N]. How much time does it take for the disk to stop?

To solve this kinematics problem, we're looking for the angular acceleration produced by the brake. We can find it with the equation $\mathcal{T} = I\alpha$. We need to find $\mathcal{T}$ and I_{disk} and solve for α. The torque produced by the brake is calculated using the force-times-leverage formula: $\mathcal{T} = F_{\perp}r = 60 \times 0.3 = 18$[N m]. The moment of inertia of a disk is given by $I_{\text{disk}} = \frac{1}{2}mR^2 = \frac{1}{2}(20)(0.3)^2 = 0.9$[kg m^2]. Thus we have $\alpha = 20$[rad/s^2]. Now we can use the UAM formula for angular velocity $\omega(t) = \alpha t + \omega_i$ and solve for the time when the object's motion will stop: $0 = \alpha t + \omega_i$. The disk will come to a stop after $t = \omega_i/\alpha = 1$[s].

Combined motion

A pulley of radius R and moment of inertia I has a rope wound around it. At the end of the rope is attached a rock of mass m. What will be the angular acceleration of the pulley if we let the rock drop to the ground while unwinding the rope?

A force diagram of the rock tells us that $mg - T = ma_y$ (where $\hat{y}$ points downward). A torque diagram of the disk tells us that $TR = I\alpha$. Taking the product of R times the first equation and adding it to the second equation gives us

$$R(mg - T) + TR = Rma_y + I\alpha,$$

and after simplification we're left with

$$Rmg = Rma_y + I\alpha.$$

Additionally, since we know the rope forms a solid connection between the pulley and the rock, this means that the angular acceleration of the pulley is related to the linear acceleration of the rock: $R\alpha = a_y$. We can use this relationship between the variables a_y and α to obtain an equation with only one unknown. We substitute $R\alpha$ for a_y in the above equation to obtain

$$Rmg = Rm(R\alpha) + I\alpha = (R^2m + I)\alpha.$$

Solving for α we find

$$\alpha = \frac{Rmg}{R^2m + I}\,.$$

This answer makes sense intuitively. From the rotating disk's point of view, the cause of rotation is the torque produced by the falling mass, while the denominator represents the total moment of inertia for the mass-pulley system as a whole.

The vertical acceleration of the falling mass is obtained via $a_y = R\alpha$:

$$a_y = \frac{R^2mg}{R^2m + I} = \frac{mg}{m + \frac{I}{R^2}}\,.$$

From the point of view of the falling mass, the cause of the motion is the weight of the object $W = mg$, while the denominator represents the total inertia of the system. Recall that *inertia* in this case refers to the notion of "resistance to motion." The effective inertia of the system is the combination of the mass m and the moment of inertia of the disk I divided by R^2.

Conservation of angular momentum

A spinning figure skater starts from an initial angular velocity of $\omega_i = 12$[rad/s] with her arms extending away from her body. In this position, her body's moment of inertia is $I_i = 3$[kg m^2]. The skater then brings her arms close to her body, and in the process her moment of inertia changes to $I_f = 0.5$[kg m^2]. What is her new angular velocity?

This looks like a job for the law of conservation of angular momentum:

$$L_i = L_f \qquad \Rightarrow \qquad I_i\omega_i = I_f\omega_f.$$

We know I_i, ω_i, and I_f, so we can solve for the final angular velocity ω_f. The answer is $\omega_f = I_i\omega_i/I_f = 3 \times 12/0.5 = 72$[rad/s], which corresponds to 11.46 turns per second.

Conservation of energy

You have a 14[in] bicycle wheel with mass $m = 4$[kg], with nearly all of its mass concentrated near the outside rim. The wheel is set in rolling motion up an incline at a velocity of 20[m/s]. How far up the incline will the wheel reach before it stops?

We can solve this problem with the principle of conservation of energy $\sum E_i = \sum E_f$. We must account for both the linear and rotational kinetic energies of the wheel:

$$\begin{aligned} K_i \quad + \quad K_{ri} \ + U_i &= K_f + K_{rf} + U_f \ , \\ \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2 + 0 \ &= \ 0 \ + \ 0 \ + mgh. \end{aligned}$$

First, calculate $I_{\rm wheel}$ using the formula $I_{\rm ring} = mR^2 = 4 \times (0.355)^2 = 0.5$[kg m^2]. If the wheel's linear velocity is 20[m/s], then its angular velocity is $\omega = v_t/R = 20/0.355 = 56.34$[rad/s]. We can now use these values in the conservation of energy equation:

$$\frac{1}{2}(4)(20)^2 + \frac{1}{2}(0.5)(56.34)^2 + 0 = 800.0 + 793.55 = (4)(9.81)h.$$

The wheel will reach a maximum height of $h = 40.61$[m].

Note that roughly half the wheel's kinetic energy is stored in its rotational motion. This demonstrates the importance of accounting for K_r when solving energy problems involving rotating objects.

Static equilibrium

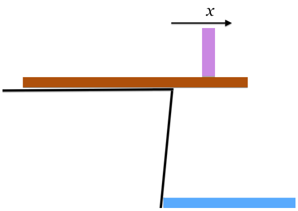
We say a system is in *equilibrium* when all forces and torques acting on the system "balance each other out." If an object is not moving, we say the object is in *static equilibrium*. Basically, zero net force implies zero motion.

We can also use this reasoning in reverse. If you see an object that is completely still, then the forces and torques acting on it must be in equilibrium:

$$\sum F_x = 0, \quad \sum F_y = 0, \quad \sum \mathcal{T} = 0.$$

Equilibrium means there is zero net force in the x-direction, zero net force in the y-direction, and zero net torque acting on the object.

Example: Walking the plank



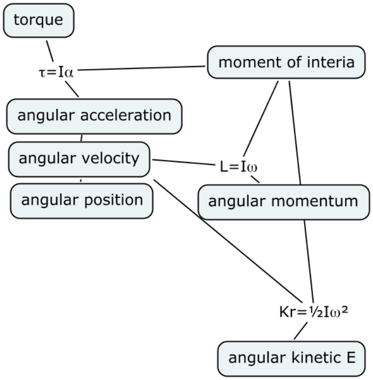
A heavy wooden plank is placed so one-third of its length protrudes from the side of a pirate ship. The plank has a total length of 12[m] and a total weight of 120[kg]; this means 40[kg] of its weight is suspended above the ocean, while 80[kg] rests on the ship's deck. How far onto the plank can a person weighing 100[kg] walk before the plank tips into the ocean?

We'll use the torque equilibrium equation $\sum \mathcal{T}_E = 0$, where we calculate the torques relative to the edge of the ship, the point around which the plank will pivot. There are two torques involved: the torque produced by the plank's weight and the torque produced by the person's weight. The plank's weight acts in its centre of mass, which is located 2[m] from the edge of the ship. The torque produced by the weight of the plank is therefore given by $\mathcal{T}_2 = 120g \times 2 = 240g$[N m]. The torque produced by the person when he reaches a distance of x[m] from the edge of the ship is $\mathcal{T}_1 = -100gx$[N m]. Thus, the maximum distance the person can walk before the plank tips is $x = \frac{240g}{100g} = 2.4$[m]. After that it's all for the sharks.

Discussion

In this section, we applied the techniques and ideas from linear motion in order to describe the rotational motion of objects. Our coverage of rotational motion has been relatively brief because there were no new notions of physics to be learned.

Calling upon our prior knowledge of physics, we explored the parallels between the new rotational concepts and their linear counterparts. It is important you understand these parallels. To help you connect the notions of rotational motion with the notions of linear motion, you can revisit the diagram on page 229 in the beginning of this section.



Let's summarize. If you know the torque acting on an object, and the object's moment of inertia, you can calculate its angular acceleration α. If you know the object's angular acceleration $\alpha(t)$, its initial angular position θ_i, and its initial angular velocity ω_i, you can calculate its equation of motion $\theta(t)$, which tells you the object's angular position at all times.

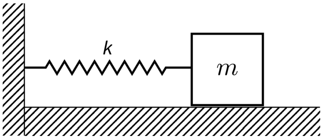
Furthermore, a rotating object's angular velocity ω is related to its *angular momentum* $L = I\omega$ and its *rotational kinetic energy* $K_r = \frac{1}{2}I\omega^2$. Angular momentum measures the "quantity of rotational motion," while rotational kinetic energy measures how much energy the object has by virtue of its rotational motion.

In rotational equations, the moment of inertia I plays the role of the mass m. In the equation $\mathcal{T} = I\alpha$, the moment of inertia I measures how difficult it is to make the object turn. The moment of inertia also appears in the formulas for finding a spinning object's angular momentum and its rotational kinetic energy.

4.9 Simple harmonic motion

Vibrations, oscillations, and waves are everywhere around us. For example, what appears to our eyes as white light is actually made of many different oscillations of the electromagnetic field. These oscillations vibrate at a range of frequencies, which correspond to the colours we perceive. Sounds are also made of combined air vibrations with various frequencies and strengths. In this section, we'll learn about *simple harmonic motion*, which describes the oscillation of a mechanical system at a fixed frequency and with a constant amplitude. As its name suggests, simple harmonic motion is the simplest form of oscillatory motion. By studying oscillations in their simplest form, you'll gather important intuition that applies to all types of oscillations and wave phenomena.

The canonical example of simple harmonic motion is the motion of a mass-spring system, illustrated in the figure. The block is free to slide along the horizontal frictionless surface. If the system is disturbed from its equilibrium position, it will start to oscillate back and forth at a certain *natural* frequency, which depends on the mass of the block and the stiffness of the spring.



We'll focus our attention on two mechanical systems: the mass-spring system and the simple pendulum. We'll follow the usual approach by describing the positions, velocities, accelerations, and energies associated with these two kinds of motion. The notion of *simple harmonic motion* (SHM) reaches farther than these two systems. The equations and intuition developed while analyzing the oscillations within these simple mechanical systems can be applied more generally to sound oscillations, electric current oscillations, and even quantum oscillations. Pay attention, is all I'm saying.

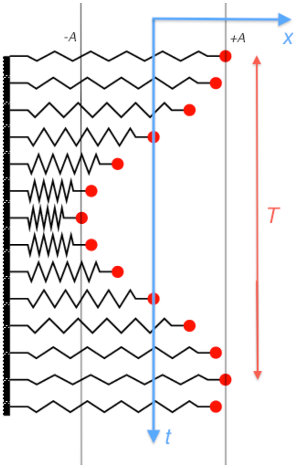
Concepts

- A: The *amplitude* of the movement is how far the object moves back and forth relative to its centre position.

- $x(t)$[m], $v(t)$[m/s], $a(t)$[m/s^2]: position, velocity, and acceleration of the object as functions of time
- T[s]: the *period* of the object's motion. The period is how long it takes for the motion to repeat.
- f[Hz]: the *frequency* of the motion
- ω[rad/s]: *angular frequency*
- ϕ[rad]: the phase shift denoted by the Greek letter *phee*

Simple harmonic motion

This figure illustrates a mass-spring system undergoing simple harmonic motion. The position of the mass as a function of time oscillates like the cosine function. From the diagram, we can identify two important parameters of the system's motion: the amplitude A, which describes the maximum displacement of the mass from the centre position, and the period T, which describes how long it takes the mass to return to its initial position.



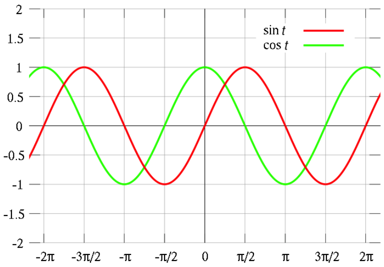
The equation that describes the object's position as a function of time is

$$x(t) = A\cos(\omega t + \phi).$$

The constant ω (omega) represents the *angular frequency* of the motion. Angular frequency is related to the period T by the equation $\omega = \frac{2\pi}{T}$. The additive constant ϕ (*phee*) is called the *phase shift*. Its value depends on the initial condition for the motion $x_i \equiv x(0)$.

I don't want you to be scared by the formula for simple harmonic motion. I know there are a lot of Greek letters in there, but it is actually pretty simple. In order to understand the purpose of the three parameters A, ω, and ϕ, let's quickly review the properties of the $\cos$ function.

Review of sin and cos functions



The functions $f(t) = \sin(t)$ and $f(t) = \cos(t)$ are periodic functions that oscillate between -1 and 1 with a period of 2π. Previously we used the functions $\cos$ and $\sin$ to find the horizontal and vertical components of vectors, and we called the input variable θ (theta). However, in this section the input variable is the time t measured in seconds. Look carefully at the plot of the function $\cos(t)$. As t goes from $t = 0$ to $t = 2\pi$, the function $\cos(t)$ completes one full cycle. We say the *period* of $\cos(t)$ is $T = 2\pi$.

Input-scaling

If we want to describe a periodic motion with a different period, we can still use the $\cos$ function, but inside the $\cos$ function we must include a multiplier before the variable t. This multiplier describes the *angular frequency* and is denoted ω (*omega*). The input-scaled $\cos$ function

$$f(t) = \cos(\omega t)$$

has a period of $T = \frac{2\pi}{\omega}$.

Scaling the input of the cos function by the constant $\omega = \frac{2\pi}{T}$ produces a periodic function with period T. When you vary t from 0 to T, the quantity ωt goes from 0 to 2π, so the function $\cos(\omega t)$ completes one cycle. You shouldn't just take my word for this; try it yourself by building a cos function with a period of 3 units.

The *frequency* of periodic motion describes the number of times per second the motion repeats. A motion's frequency is equal to the inverse of its period:

$$f = \frac{1}{T} = \frac{\omega}{2\pi} \text{ [Hz]}.$$

Frequency f and angular frequency ω are related by a factor of 2π. We need this multiplier since the natural cycle length of the $\cos$ function is 2π radians.

Output-scaling

We can scale the output of the $\cos$ function by a constant A, called the *amplitude*. The function

$$f(t) = A\cos(\omega t)$$

will oscillate between $-A$ and A.

Time-shifting

The motion described by the function $A\cos(\omega t)$ starts from its maximum value at $t = 0$. A mass-spring system described by the position function $x(t) = A\cos(\omega t)$ begins its motion with the spring maximally stretched $x_i \equiv x(0) = A$.

If we want to describe other starting positions for the motion, it may be necessary to introduce a *phase shift* inside the $\cos$ function:

$$x(t) = A\cos(\omega t + \phi).$$

The constant ϕ must be chosen so that at $t = 0$, the function $x(t)$ correctly describes the initial position of the system.

For example, if the harmonic motion starts from the system's centre $x_i \equiv x(0) = 0$ and initially moves in the positive direction, then the motion is described by the function $A\sin(\omega t)$. Or, since $\sin(\theta) = \cos(\theta - \frac{\pi}{2})$, we can describe the same motion in terms of a shifted $\cos$ function:

$$x(t) = A\cos\Big(\omega t - \frac{\pi}{2}\Big) = A\sin(\omega t).$$

Note, the function $x(t)$ correctly describes the system's initial position $x(0) = 0$.

The constant in front of the $\cos$ tells us the motion's amplitude A, and the multiplicative constant ω inside the $\cos$ is related to the motion's period/frequency: $\omega = \frac{2\pi}{T} = 2\pi f$. Finally, the additive constant ϕ is chosen depending on the initial conditions. By now, the meaning of all the parameters in the simple harmonic motion equation should be clear to you.

Mass and spring

Okay, it's time to apply all this math to a physical system which exhibits simple harmonic motion: the mass-spring system.

An object of mass m is attached to a spring with spring constant k. If disturbed from rest, this mass-spring system will undergo simple harmonic motion with angular frequency of

$$\omega = \sqrt{\frac{k}{m}}\,.$$

A stiff spring attached to a small mass will result in very rapid oscillations. A weak spring or a heavy mass will result in slow oscillations.

A typical exam question may tell you k and m and ask about the period T. If you remember the definition of T, you can easily calculate the answer:

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{m}{k}}.$$

Equations of motion

The general equations of motion for the mass-spring system are

$$x(t) = A\cos(\omega t + \phi), \tag{4.4}$$

$$v(t) = -A\omega\sin(\omega t + \phi), \tag{4.5}$$

$$a(t) = -A\omega^2\cos(\omega t + \phi). \tag{4.6}$$

The general shape of the function $x(t)$ is similar to that of a $\cos$ function. The *angular frequency* ω parameter is governed by the physical properties of the system. The parameters A and ϕ describe the specifics of the motion, namely, the *amplitude* of the oscillation and its starting position.

The function $v(t)$ is obtained, as usual, by taking the derivative of $x(t)$. The function $a(t)$ is obtained by taking the derivative of $v(t)$, which corresponds to the second derivative of $x(t)$. The velocity and acceleration are also periodic functions.

Motion parameters

The key motion parameter of SHM is how far the mass swings back and forth through the *centre position*. The amplitude A describes the maximum distance the mass will travel in the positive x-direction. We can also find the maximum values of an object's velocity and acceleration by reading the coefficient located in front of $\sin$ and $\cos$ in the functions $v(t)$ and $a(t)$.

- The object's maximum velocity is $v_{\max} = A\omega$.
- The object's maximum acceleration is $a_{\max} = A\omega^2$.

The velocity function reaches its maximum as the object passes through the centre position. The acceleration is maximum when the spring is maximally stretched or compressed—these are the locations where the pull of the spring is the strongest.

You'll definitely be asked to solve for the quantities $v_{\max}$ and $a_{\max}$ in exercises and exams. This is an easy task if you remember the above formulas and you know the values of the amplitude A and the angular frequency ω.

Energy

The potential energy stored in a spring that is stretched or compressed by a length x is given by the formula $U_s = \frac{1}{2}kx^2$. Since we know $x(t)$, we can obtain the potential energy of the mass-spring system as a function of time:

$$U_s(t) = \frac{1}{2}kx(t)^2 = \frac{1}{2}kA^2\cos^2(\omega t + \phi).$$

The potential energy reaches its maximum value $U_{s,\max} = \frac{1}{2}kA^2$ when the spring is fully stretched or fully compressed.

The kinetic energy of the mass as a function of time is given by

$$K(t) = \frac{1}{2}mv(t)^2 = \frac{1}{2}m\omega^2A^2\sin^2(\omega t + \phi).$$

The kinetic energy is maximum when the mass passes through the centre position. The maximum kinetic energy is given by $K_{\max} = \frac{1}{2}mv_{\max}^2 = \frac{1}{2}mA^2\omega^2$.

Conservation of energy

Since the mass-spring system does not experience any dissipative forces (friction), the total energy of the system $E_T(t) = U_s(t) + K(t)$ must be *conserved*. The conservation of energy principle says that the sum of the system's potential energy and its kinetic energy must be the same at any two instants t_1 and t_2:

$$E_T(t_1) = U_s(t_1) + K(t_1) = U_s(t_2) + K(t_2) = E_T(t_2).$$

According to this equation, even if $U_s(t)$ and $K(t)$ change over time, the system's total energy $E_T(t)$ always remains constant.

Let us convince ourselves the system's total energy is indeed a constant. We can use the identity $\cos^2\theta + \sin^2\theta = 1$ to find the value of this constant:

$$\begin{aligned}
E_T(t) &= U_s(t) + K(t) \\
&= \frac{1}{2}kA^2\cos^2(\omega t) + \frac{1}{2}m\omega^2A^2\sin^2(\omega t) \\
&= \frac{1}{2}m\omega^2A^2\cos^2(\omega t) + \frac{1}{2}m\omega^2A^2\sin^2(\omega t) \quad \text{(since } k = m\omega^2\text{)} \\
&= \frac{1}{2}m\underbrace{\omega^2A^2}_{v_{\max}^2}\underbrace{\left[\cos^2(\omega t) + \sin^2(\omega t)\right]}_{=1} = \frac{1}{2}mv_{\max}^2 = K_{\max} \\
&= \frac{1}{2}m(\omega A)^2 = \frac{1}{2}(m\omega^2)A^2 = \frac{1}{2}kA^2 = U_{s,\max}.
\end{aligned}$$

The system's total energy is equal to $U_{s,\max}$ and to $K_{\max}$.

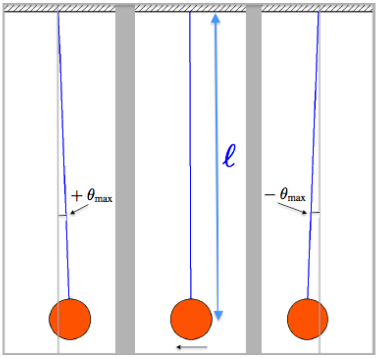
The best way to understand simple harmonic motion is to visualize how the system's energy shifts between the spring's potential energy and the kinetic energy of the moving mass. When the spring is maximally stretched $x = \pm A$, the mass will have zero velocity and thus zero kinetic energy $K = 0$. At this moment of maximal displacement, all the system's energy is stored in the potential energy of the spring $E_T = U_{s,\max}$.

The other important moment happens when the mass has zero displacement. In this moment the position $x = 0$ implies there is zero potential energy in the spring $U_s = 0$. This is when the velocity is maximum $v = \pm A\omega$ and all the system's energy is stored entirely in its kinetic energy $E_T = K_{\max}$.

Pendulum motion

We now turn our attention to another simple mechanical system in which motion is described by the simple harmonic motion equations: the pendulum.

A pendulum is a mass suspended at the end of a string of length ℓ. Imagine pulling the mass so it is positioned a certain angle $\theta_{\max}$ away from the pendulum's vertical resting position. When you release the mass, the pendulum swings back and forth undergoing simple harmonic motion.



In a gravitational field of strength g, the pendulum's period of oscillation is given by the formula

$$T = 2\pi\sqrt{\frac{\ell}{g}}.$$

Note, the period does not depend on the amplitude of the oscillation (how far the pendulum swings), nor does the period depend on the pendulum's mass. The only factors that play a role in determining the period of oscillation are the length of the string ℓ and the strength of the gravitational field g. Recall that angular frequency is defined as $\omega = \frac{2\pi}{T}$, so the angular frequency for the pendulum is

$$\omega \equiv \frac{2\pi}{T} = \sqrt{\frac{g}{\ell}}.$$

Instead of describing the pendulum's position x with respect to Cartesian coordinate system, we describe its position in terms of the angle θ it makes with the vertical line that passes through the centre of the motion. The equations of motion are described in terms of *angular quantities*: the angular position θ, the angular velocity ω_θ, and the angular acceleration α_θ of the pendulum:

$$\theta(t) = \theta_{\max}\,\cos\left(\sqrt{\frac{g}{\ell}}t + \phi\right),$$

$$\omega_\theta(t) = -\theta_{\max}\sqrt{\frac{g}{\ell}}\,\sin\!\left(\sqrt{\frac{g}{\ell}}t + \phi\right),$$
$$\alpha_\theta(t) = -\theta_{\max}\frac{g}{\ell}\,\cos\!\left(\sqrt{\frac{g}{\ell}}t + \phi\right).$$

The angle $\theta_{\max}$ describes the maximum angle to switch the pendulum swings. Notice the new variable name ω_θ we use for the pendulum's angular velocity $\omega_\theta(t) = \frac{d}{dt}(\theta(t))$. The angular velocity ω_θ of the pendulum should not be confused with the *angular frequency* $\omega = \sqrt{\frac{g}{\ell}}$ of the periodic motion, which is the constant inside the $\cos$ function.

Energy

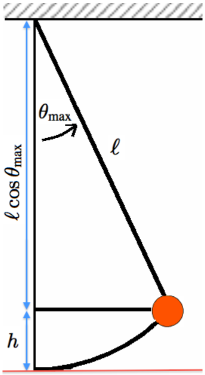
A pendulum's motion is best understood by imagining how the energy in the system shifts between gravitational potential energy and kinetic energy.

The pendulum reaches its maximum potential energy when it swings sideways to reach angle $\theta_{\max}$. At this angle, the mass's vertical position is increased by a height h above the mass's lowest point. We can calculate h as follows:

$$h = \ell - \ell\cos\theta_{\max}.$$

The maximum gravitational potential energy of the mass is therefore

$$U_{g,\max} = mgh = mg\ell(1 - \cos\theta_{\max}).$$



By the conservation of energy principle, the pendulum's maximum kinetic energy must equal its maximum gravitational potential energy:

$$mg\ell(1 - \cos\theta_{\max}) = U_{g,\max} = K_{\max} = \frac{1}{2}mv_{\max}^2,$$

where $v_{\max} = \ell\omega_\theta$ is the linear velocity of the mass as it swings through the vertical position.

Explanations

It's worthwhile to understand where the simple harmonic motion equation comes from. In this subsection, we'll discuss how the equation $x(t) = A\cos(\omega t + \phi)$ is derived from Newton's second law $F = ma$ and the equation for the force of a spring $F_s = -kx$.

Trigonometric derivatives

The slope (derivative) of the function $\sin(t)$ varies between -1 and 1. The slope is largest when $\sin$ passes through the x-axis, and the slope is zero when the function reaches its maximum and minimum values. A careful examination of the graphs of the bare functions $\sin$ and $\cos$ reveals that the derivative of the function $\sin(t)$ is described by the function $\cos(t)$, and vice versa:

$$f(t) = \sin(t) \qquad \Rightarrow \qquad f'(t) = \cos(t),$$
$$f(t) = \cos(t) \qquad \Rightarrow \qquad f'(t) = -\sin(t).$$

When you learn more about calculus, you'll know how to find the derivative of any function you want; for now, you can take my word that the above two formulas are true.

The chain rule for derivatives dictates that a composite function $f(g(x))$ has derivative $f'(g(x)) \cdot g'(x)$. First we take the derivative of the outer function, then we multiply by the derivative of the inner function. We can find the derivative of the position function $x(t) = A\cos(\omega t + \phi)$ using the chain rule:

$$v(t) \equiv x'(t) = -A\sin(\omega t + \phi) \cdot \omega = -A\omega \sin(\omega t + \phi),$$

where the outer function is $f(x) = A\cos(x)$ with derivative $f'(x) = -A\sin(x)$, and the inner function is $g(x) = \omega x + \phi$ with derivative $g'(x) = \omega$.

The same reasoning is applied to obtain the second derivative:

$$a(t) \equiv \frac{d}{dt}\{v(t)\} = -A\omega^2 \cos(\omega t + \phi) = -\omega^2 x(t).$$

Note the function $a(t) \equiv x''(t)$ has the same form as the function $x(t)$; the two functions differ only by the factor $-\omega^2$.

Derivation of the mass-spring SHM equation

You may be wondering where the equation $x(t) = A\cos(\omega t + \phi)$ comes from. This formula looks very different from the kinematics equation for linear motion $x(t) = x_i + v_i t + \frac{1}{2}at^2$, which we obtained starting with Newton's second law $F = ma$ and completing two steps of integration.

In this section, I've seemingly pulled the $x(t) = A\cos(\omega t + \phi)$ formula out of thin air, as if by revelation. Why did we suddenly start talking about $\cos$ functions and Greek letters with dubious names like "phase"? Are you phased by all of this? When I was first learning about simple harmonic motion, I was totally phased because I didn't see where the $\sin$ and $\cos$ were coming from.

The $\cos$ also comes from $F = ma$, but the story is a little more complicated this time. The force exerted by a spring is $F_s = -kx$. Since we assume the surface the mass slides along is frictionless, the only force acting on the mass is the force of the spring:

$$\sum F = F_s = ma \qquad \Rightarrow \qquad -kx = ma.$$

Recall that the acceleration function is the second derivative of the position function:

$$a = \frac{dv(t)}{dt} = \frac{d^2x(t)}{dt^2} = x''(t).$$

We can rewrite the equation $-kx = ma$ in terms of the function $x(t)$ and its second derivative:

$$\begin{aligned} -kx(t) &= m\frac{d^2x(t)}{dt^2} \\ -\frac{k}{m}x(t) &= \frac{d^2x(t)}{dt^2}, \end{aligned}$$

which can be rewritten as

$$0 = \frac{d^2x(t)}{dt^2} + \frac{k}{m}x(t).$$

This is called a *differential equation*. Instead of looking for an *unknown number* as in normal equations, in differential equations we are looking for an *unknown function* $x(t)$. We do not know what $x(t)$ is, but we do know one of its properties—namely, that $x(t)$'s second derivative $x''(t)$ is equal to the negative of $x(t)$ multiplied by some constant.

To solve a differential equation, you must guess which function $x(t)$ satisfies this property. There is an entire course called Differential Equations, in which engineers and physicists learn how to do this guessing thing. Can you think of a function that, when multiplied by $\frac{k}{m}$, is equal to its second derivative?

Okay, I thought of one:

$$x_1(t) = A_1 \cos\left(\sqrt{\frac{k}{m}} t\right).$$

Come to think of it, there is also a second function that works:

$$x_2(t) = A_2 \sin\left(\sqrt{\frac{k}{m}} t\right).$$

You should try this for yourself. Verify that $x_1''(t) + \frac{k}{m} x_1(t) = 0$ and $x_2''(t) + \frac{k}{m} x_2(t) = 0$, which means these functions are *both* solutions to the differential equation $x''(t) + \frac{k}{m} x(t) = 0$. Since both $x_1(t)$ and $x_2(t)$ are solutions, any combination of them must also be a solution:

$$x(t) = A_1 \cos(\omega t) + A_2 \sin(\omega t).$$

This is *kind of* the answer we're looking for: an expression that describes the object's position as a function of time. I say *kind of* because the solution we obtained is not specified as a $\cos$ function with amplitude A and a phase ϕ, but instead in terms of the coefficients A_1 and A_2, which describe the $\cos$ and $\sin$ components of the motion.

Lo and behold, using the trigonometric identity from page 88 $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$, we can rewrite the above expression for $x(t)$ as a time-shifted trigonometric function:

$$x(t) = A\cos(\omega t + \phi) = A_1 \cos(\omega t) + A_2 \sin(\omega t).$$

The expression on the left is the preferred way of describing simple harmonic motion because the parameters A and ϕ correspond to observable aspects of the motion.

Let's review one more time: we are looking for the equation of motion that predicts an object's position as a function of time $x(t)$. We can draw an analogy to a situation we've seen before. In linear kinematics, uniform accelerated motion with $a(t) = a$ is described by the equation $x(t) = x_i + v_i t + \frac{1}{2}at^2$ in terms of parameters x_i and v_i. Depending on the object's initial position and initial velocity, we obtain different trajectories. Simple harmonic motion with angular frequency ω is described by the equation $x(t) = A\cos(\omega t + \phi)$ in terms of the parameters A and ϕ. Depending on the values of the amplitude A and the phase ϕ, we obtain different simple harmonic motion trajectories.

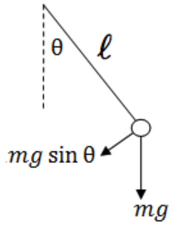
Derivation of the pendulum SHM equation

To see how the simple harmonic motion equation for the pendulum is derived, we need to start from the torque equation $\mathcal{T} = I\alpha$.

The diagram illustrates how we can calculate the torque on the pendulum, which is caused by the force of gravity as a function of the mass's displacement angle θ. Recall the torque calculation only accounts for the $F_\perp$ component of any force, since this is the only part of the force that causes rotation:

$$\mathcal{T}_\theta = F_\perp \ell = -mg\sin\theta\ell.$$

The torque is negative because it acts in the opposite direction to the displacement angle θ.



Now we substitute this expression for $\mathcal{T}_\theta$ into the angular version of Newton's second law $\mathcal{T} = I\alpha$ to obtain

$$\begin{aligned} \mathcal{T} &= I\alpha \\ -mg\sin\theta(t)\ell &= m\ell^2 \frac{d^2\theta(t)}{dt^2} \\ -g\sin\theta(t) &= \ell\frac{d^2\theta(t)}{dt^2}\,. \end{aligned}$$

To continue with the derivation, we must make an approximation. When θ is a small angle, we can use the following approximation:

$$\sin(\theta) \;\approx\; \theta, \qquad\qquad \text{for } \theta \ll 1.$$

This is known as a *small angle approximation*. You'll see where it comes from later when you learn about Taylor series approximations to functions. For now, you can convince yourself of the above formula by zooming in near the origin on the graph of the function $\sin x$ until you realize $y = \sin(x)$ looks very much like $y = x$.

Using the small angle approximation $\sin\theta \approx \theta$, we rewrite the equation involving $\theta(t)$ and its second derivative as

$$\begin{aligned} -g\sin\theta(t) &= \ell\frac{d^2\theta(t)}{dt^2} \\ -g\theta(t) &\approx \ell\frac{d^2\theta(t)}{dt^2} \\ 0 &= \frac{d^2\theta(t)}{dt^2} + \frac{g}{\ell}\theta(t). \end{aligned}$$

Now we can recognize that we're dealing with the same differential equation as in the case of the mass-spring system: $\theta''(t) + \omega^2\theta(t) = 0$, which has the solution

$$\theta(t) = \theta_{\max}\cos(\omega t + \phi),$$

where the constant inside the $\cos$ function is $\omega = \sqrt{\frac{g}{\ell}}$.

Examples

Most word problems will usually tell you the initial amplitude $x_i = A$ or the initial velocity $v_i = \omega A$ of the SHM and ask you to calculate some other quantity. Answering these problems shouldn't be too difficult if you write down the general equations for $x(t)$, $v(t)$, and $a(t)$, fill-in the known quantities, and then solve for the unknowns.

Example

You are observing a mass-spring system built from a 1[kg] mass and a 250[N/m] spring. The amplitude of the mass's oscillation is 10[cm]. Determine (a) the mass's maximum speed, (b) the maximum acceleration, and (c) the total mechanical energy in the system.

First we must find this system's angular frequency: $\omega = \sqrt{k/m} = \sqrt{250/1} = 15.81$[rad/s]. To find (a) we use the equation $v_{\max} = \omega A = 15.81 \times 0.1 = 1.58$[m/s]. Similarly, we can find the maximum acceleration using $a_{\max} = \omega^2 A = 15.81^2 \times 0.1 = 25$[m/s^2]. There are two equivalent ways to solve for (c). We can obtain the system's total energy by considering the spring's potential energy when it is maximally stretched or maximally compressed: $E_T = U_s(A) = \frac{1}{2}kA^2 = 1.25$[J]. Or, we can obtain the total energy from the maximum kinetic energy $E_T = K = \frac{1}{2}mv_{\max}^2 = 1.25$[J].

Discussion

In this section we learned about simple harmonic motion, which is described by the equation $x(t) = A\cos(\omega t + \phi)$. You may be wondering what *non-simple* harmonic motion is. You could extend what we've learned by studying oscillating systems where the energy is slowly dissipating. This is known as *damped harmonic motion*, for which the equation of motion looks like $x(t) = Ae^{-\gamma t}\cos(\omega t+\phi)$. This equation describes an oscillation with an amplitude that slowly decreases. The coefficient γ is known as the damping coefficient, and indicates how quickly the system's energy dissipates.

The concept of simple harmonic motion arises in many other areas of physics. When you learn about electric circuits, capacitors, and inductors, you'll run into equations of the form $v''(t) + \omega^2 v(t) = 0$, which indicates that a circuit's *voltage* is undergoing simple harmonic motion. Guess what—the same equation that describes the mechanical motion of the mass-spring system is used to describe the voltage in an oscillating circuit!

Links

[Plot of the simple harmonic motion using a can of spray-paint]
http://www.youtube.com/watch?v=p9uhmjbZn-c

[15 pendulums with different lengths]
http://www.youtube.com/watch?v=yVkdfJ9PkRQ

4.10 Conclusion

The fundamental purpose of mechanics is to predict the motion of objects using equations. In the beginning of the chapter, I claimed there are only 20 equations you need to know in order to solve any physics problem. Let us verify this claim and review the material we've covered.

Our goal was to find $x(t)$ for all times t. However, none of the equations of physics tell us $x(t)$ directly. Instead, we have Newton's second law $F = ma$, which tells us that the acceleration of the object $a(t)$ equals the *net force* acting on the object divided by the object's mass. To find $x(t)$ starting from $a(t)$, we use integration twice:

$$\frac{1}{m}\left(\sum \vec{F} \equiv \vec{F}_{\text{net}}\right) = a(t) \overset{v_i+\int dt}{\longrightarrow} v(t) \overset{x_i+\int dt}{\longrightarrow} x(t).$$

We studied kinematics in several different contexts. We originally looked at kinematics problems in one dimension, and derived the UAM and UVM equations. We also studied the problem of projectile motion by deconstructing it into two separate kinematics subproblems: one in the x-direction (UVM), and one in the y-direction (UAM). Later, we studied the circular motion of objects and stated equation $a_r = \frac{v_t^2}{r}$, which describes an important relationship between the radial acceleration, the tangential velocity, and the radius of the circle of rotation. We also studied rotational motion using angular kinematics quantities $\theta(t)$, $\omega(t)$, and $\alpha(t)$. We defined the concept of *torque* and saw the role it plays in the angular equivalent of Newton's second law $\mathcal{T} = I\alpha$. We studied the equation that describes simple harmonic motion, $x(t) = A\cos(\omega t + \phi)$, and showed the formula $\omega = \sqrt{\frac{k}{m}}$, which gives the angular frequency of a mass-spring system.

We also discussed three conservation laws: the conservation of linear momentum law $\sum \vec{p}_i = \sum \vec{p}_f$, the conservation of angular momentum law $L_i = L_f$, and the conservation of energy law $\sum E_i = \sum E_f$. Each of these three fundamental quantities is conserved overall and can neither be created nor destroyed. Momentum calculations are used to analyze collisions, while energy formulas like equations $K = \frac{1}{2}mv^2$, $U_g = mgh$, and $U_s = \frac{1}{2}kx^2$ can be used to analyze the motion of objects in terms of energy principles.

Now you can see how 20 equations truly are enough to master all of mechanics. Nice work! Your next step should be to practice solving exercises in order to solidify your understanding.

4.11 Mechanics exercises

You can test your understanding with the physics exercises presented in this section. Don't be discouraged if you find the exercises difficult—these are meant to be hard.

When solving exercises, I recommend you follow these steps:

1. Figure out what type of problem you are dealing with. Kinematics? Angular motion? Energy?
2. Draw a diagram that describes the physical situation. Label things clearly.
3. Copy from your formula sheet all the equations you plan to use.
4. Substitute the known quantities into the equations, and determine which unknown you are looking for. Visualize the steps you will use to solve for the unknowns.
5. Solve for the unknowns.

Make sure you attempt each of the exercises on your own before looking at the solutions.

Simple ones

Simple kinematics

A ball is thrown from the ground at an initial upward velocity of 20[m/s]. How long will the ball stay in the air before it returns to the ground?
Sol: This is a kinematics question. Start from the equation $v(t) = at + v_i$ and $a = -9.81$. We know $v(t_{\text{top}}) = 0$, so we can solve this equation to find t_{top}. Ans: $t_{\text{flight}} = 2t_{\text{top}} = 4.1$[s].

More kinematics

Given $a(t) = 4$[m/s^2], $v_i = 10$[m/s], $x_i = 20$[m], find $x(t)$, the position as a function of time t[s].
Ans: $x(t) = 2t^2 + 10t + 20$.

Good ones

Turntable slug

A disk is rotating with an angular velocity of $\omega = 5$[rad/s]. A slug is sliding along the surface of the disk in the radial direction. The slug started from the disk's centre and has been moving outward. If the coefficient of friction between the slug and the disk is $\mu_k = 0.4$, how far can the slug slide before it flies off the surface of the disk?
Ans: The normal force between the slug and the turntable is $N = mg$. With the slug located at radius R, the centripetal acceleration required to keep the slug on the disk is $F_r = ma_r = m\frac{(R\omega)^2}{R}$. The friction force available is $F_f = 0.4mg$. The slug will fly off when the friction force becomes insufficient. The slug will lose its grip at a distance $R = \frac{0.4g}{\omega^2}$ from the centre of the disk.

Word problems

Elevator fridge

You have loaded a fridge into an elevator. Due to the static force of friction, the refrigerator needs a strong push to start it sliding across the elevator floor. From smallest to largest, rank the magnitude of the static force of friction in these three situations: a stationary elevator, an upward accelerating elevator, and a downward accelerating elevator.
Ans: upward $F_{fs} >$ stationary $F_{fs} >$ downward F_{fs}. The equation for F_{fs} is $F_{fs} = \mu_s N$, where N is the normal force (the contact force between the fridge and the elevator floor). In the y-direction, the force diagram on the elevator reads $\sum F_y = N - mg = ma_y$. When the elevator is static, $a_y = 0$ so $N = mg$. If $a_y > 0$ (upward acceleration), then we must have $N > mg$; hence the friction force will be larger than when the elevator is static. When $a_y < 0$ (downward acceleration), N must be smaller than mg, and consequently there will be less F_{fs}.

More turntable tricks

Three coins are placed on a turntable. One coin is placed 5[cm] from the turntable's centre, another is placed 10[cm] from the centre, and the third is placed 15[cm] from the centre. The turntable is powered on and begins to spin. Initially, due to static friction, the coins move together with the turntable as it starts rotating. The angular speed ω of the turn table then increases slowly. Assuming all the coins have the same coefficient of friction with the turntable surface, which coin begins to slide first?

Sol: This is a circular motion question. The coin furthest from the centre will be the first to fly off the spinning turntable. That is because the centripetal force required to keep this coin turning is the largest. Recall that $F_r = ma_r$, that $a_r = v^2/R$, and that $v = \omega R$. If the turntable turns with angular velocity ω, the centripetal acceleration required to keep a coin turning in a circle of radius R will be $F_r = m\omega^2 R$. This centripetal force must be supplied by the static force of friction F_{fs} between the coin and the turntable. A large R requires more F_{fs}, hence the furthest coin will fly off first.

Leverage is key

Two identical pulleys with the same moment of inertia have strings wound around them. The first pulley's string is wound around the outer radius R, while the second pulley's string is wound around a smaller radius $r < R$. The same force F is applied to rotate the pulleys. After a fixed time t, which pulley has the faster rotational speed? Which pulley has the greater rotational kinetic energy?
Sol: This is an angular motion question. The two pulleys have identical rotational resistance, meaning they have the same moment of inertia I. The string wound around the larger radius R will produce the larger torque. Higher torque will produce more angular acceleration and therefore a bigger angular velocity and kinetic energy.

Integration tests

The following exercises require you to mix techniques from different sections.

Disk brakes

The disk brake pads on your new bicycle squeeze the brake disks with a force of 5000[N] from each side. There is one brake pad on each tire. The coefficient of friction between brake pads and brake disks is $\mu_k = 0.3$. The brake disks have a radius of $r = 6$[cm], and the tire has radius $R = 20$[cm].
1. What is the total force of friction exerted by each brake?
2. What is the torque exerted by each brake?
3. You are moving at 10[m/s] and apply broth brakes. The combined mass of you and

your bicycle is 100[kg]. How many times will the wheels turn before the bike stops?
4. What will be the braking distance?
Sol: 1. Friction force is proportional to normal force, so the friction on each side of each disk is $F_f = 0.3 \times 5000 = 1500$[N], for a total friction force of $F_f = 3000$[N] per wheel.
2. The friction force of the brakes acts with a leverage of 0.06[m], so the torque produced by each brake is $\mathcal{T} = 0.06 \times 3000 = 180$[N m].
3. The kinetic energy of a 100[kg] object moving at 10[m/s] is equal to $K_i = \frac{1}{2}100(10)^2 = 5000$[J]. We'll use $K_i - W = 0$, where W is the work done by the brakes. Let θ_{stop} be the angle of rotation of the tires when the bike stops. The work done by each brake is $180\theta_{\text{stop}}$. It will take a total of $\theta_{\text{stop}} = \frac{5000}{360} = 13.\overline{8}$[rad] to stop the bike. This angle corresponds to 2.21 turns of the wheels.
4. Your stopping distance will be $13.\overline{8} \times 0.20 = 2.\overline{7}$[m]. Yay for disk brakes!

Tarzan

A half-naked dude swings from a long rope attached to the ceiling. The rope has a length of 6[m]. The dude swings from an initial angle of $-50°$ ($50°$ to the left of the rope's vertical line) all the way to the angle $+10°$, at which point he lets go of the rope. How far will Tarzan fall, as measured from the centre position of the swing motion? I am asking you to find $x_f = 6\sin(10) + d$ where d is the distance travelled by Tarzan after he lets go of the rope.
Sol: This is an energy problem followed by a projectile motion problem. The energy equation $\sum E_i = \sum E_f$ is expressed in this case as $U_i = U_f + K_f$, or $mg(6-6\cos 50°) = mg(6-6\cos 10°)+\frac{1}{2}mv^2$, which can be simplified to $v^2 = 12g(\cos 10° - \cos 50°)$. Solving for v we find $v = 4.48$[m/s]. Now for the projectile motion part. The initial velocity is 4.48[m/s] at an angle of $10°$ with respect to the ground, so $v_i = (4.42, 0.778)$[m/s]. Tarzan's initial position is $(x_i, y_i) = (6\sin(10), 6[1-\cos(10)]) = (1.04, 0.0911)$[m]. To find the total time of flight, we solve for t in $0 = -4.9t^2 + 0.778t + 0.0911$ and find $t = 0.237$[s]. Tarzan will fly for a distance of $x_f = 6\sin(10) + 4.42 \times 0.237 = 2.08$[m].

Advanced

Pendulum painting

Two disgruntled airport employees decide to vandalize a moving walkway by suspending a leaking-paint-bucket pendulum above it. The pendulum is composed of a long cable (considered massless) and a paint bucket with a hole in the bottom. The pendulum's oscillations are small, and transverse to the direction of the walkway's motion. Find the equation $y(x)$ of the resulting pattern of paint on the moving walkway in terms of the pendulum's maximum (angular) displacement $\theta_{\max}$, its length ℓ, and the speed of the walkway v. Assume x measures distances along the walkway and y denotes the pendulum's transversal displacement.
Sol: This is a simple harmonic motion question involving a pendulum. Begin by writing the general equation of motion for a pendulum: $\theta(t) = \theta_{\max}\cos(\omega t)$, where $\omega = \sqrt{g/\ell}$. Enter the walkway, which is moving to the left at velocity v. If we choose the $x = 0$ coordinate at a time when $\theta(t) = \theta_{\max}$, the pattern on the walkway can be described by the equation $y(x) = \ell\sin(\theta_{\max})\cos(kx)$, where $k = 2\pi/\lambda$, and λ tells us how long (measured as a distance in the x-direction) it takes for the pendulum to complete one cycle. One full swing of the bucket takes $T = 2\pi/\omega$[s]. In that time, the moving walkway will have moved a distance of vT metres. So one cycle in space (one wavelength) is $\lambda = vT = v2\pi/\omega$. We conclude that the equation of the paint on the moving sidewalk is $y(x) = \ell\sin(\theta_{\max})\cos((\omega/v)x)$. Observe that the angular frequency parameter $\omega = \sqrt{g/\ell}$ does not depend on the mass of the pendulum; thus the change in mass as the paint leaks will not affect the pendulum's motion.

Links

[Physics exercises]
http://en.wikibooks.org/wiki/Physics_Exercises
http://en.wikibooks.org/wiki/A-level_Physics_(Advancing_Physics)

[Lots of examples with solutions]
http://farside.ph.utexas.edu/teaching/301/lectures/lectures.html

Chapter 5

Calculus

Calculus is *useful* math. We use calculus to solve problems in physics, chemistry, computing, biology, and many other areas of science. You need calculus to perform the quantitative analysis of how functions change over time (derivatives), and to calculate the total amount of a quantity that accumulates over a time period (integrals).

The language of calculus will allow you to speak precisely about the properties of functions and better understand their behaviour. You will learn how to calculate the slopes of functions, how to find their maximum and minimum values, how to compute their integrals, and other tasks of practical importance.

5.1 Introduction

In Chapter 2, we developed an intuitive understanding of integrals. Starting with the knowledge of an object's acceleration function over time, we used the integration operation to calculate the object's velocity function and its position function. We'll now take a closer look at the techniques of calculus using precise mathematical statements, and study how these techniques apply to other problems in science.

A strong knowledge of functions is essential for your understanding of the new calculus concepts. I recommend revisiting Section 1.14 (page 58) to remind yourself of the functions introduced therein. I insist on this. Go! Seriously, there is no point in learning that the derivative of the function $\sin(x)$ is the function $\cos(x)$ if you don't have a clue what $\sin(x)$ and $\cos(x)$ are.

Before we introduce any formal definitions, formulas or derivations, let's demonstrate how calculus is used in a real-world example.

Download example

Suppose you're downloading a large file to your computer. At $t = 0$ you click "save as" in your browser and the download starts. Let $f(t)$ represent the size of the downloaded data. At any time t, the function $f(t)$ tells you the amount of disk space taken by the partially-downloaded file. You are downloading a 720[MB] file, so the download progress at time t corresponds to the fraction $\frac{f(t)}{720\text{[MB]}}$.

Download rate

The derivative function $f'(t)$, pronounced "f prime," describes how the function $f(t)$ changes over time. In our example $f'(t)$ is the download speed. If your downloading speed is $f'(t) = 100$[kB/s], then the file size $f(t)$ must increase by 100[kB] each second. If you maintain this download speed, the file size will grow at a constant rate: $f(0) = 0$[kB], $f(1) = 100$[kB], $f(2) = 200$[kB], $\ldots$, $f(100) = 10$[MB].

To calculate the "estimated time remaining" until the download's completion, we divide the amount of data that remains to be downloaded by the current download speed:

$$\text{time remaining} = \frac{720 - f(t)}{f'(t)} \quad [\text{s}].$$

The bigger the derivative, the faster the download will finish. If your Internet connection were 10 times faster, the download would finish 10 times more quickly.

Inverse problem

Let's consider this situation from the point of view of the modem that connects your computer to the Internet. Any data you download comes through the modem. The modem knows the download rate $f'(t)$[kB/s] at all times during the download.

However, since the modem is separate from your computer, it does not know the file size $f(t)$ as the download progresses. Nevertheless, the modem can infer the file size at time t from knowing the transmission rate $f'(t)$. The integral of the download rate between $t = 0$ and $t = \tau$ corresponds to the total amount of downloaded data stored on your computer. During this download period, the change in file size is described by the integral

$$\Delta f = f(\tau) - f(0) = \int_0^\tau f'(t)\, dt\,.$$

Assuming the file size starts from zero $f(0) = 0$[kB] at $t = 0$, the modem can use the integration procedure to find $f(\tau)$, the file size on your computer at $t = \tau$:

$$f(\tau) = \int_0^\tau f'(t)dt\,.$$

The download rate $f'(t)$ is measured in [kB/s], and each time step dt is 1[s] long, so the data downloaded during one second is $f'(t)dt$[kB]. The file size at time $t = \tau$ is equal to the sum of the data downloaded during each second from $t = 0$ until $t = \tau$.

The integral $\int_a^b q(t)\, dt$ is the calculation of the *total* of some quantity $q(t)$ that accumulates during the time period from $t = a$ to $t = b$. Integrals are necessary any time you want to calculate the total of a quantity that changes over time.

As demonstrated above, calculus is much more than the theoretical activity reserved for math specialists. Calculus relates to everyday notions you're already familiar with. Indeed, we carry out calculus-like operations in our head every day—we just don't necessarily use calculus terminology when we do so.

Learning the language of calculus will help you think more clearly about certain types of problems. Understanding the language of calculus is *essential* for learning science because many laws of nature are best described in terms of derivatives and integrals.

Usually, differential calculus and integral calculus are taught as two separate subjects. Perhaps teachers and university administrators are worried the undergraduates' little heads will explode from sudden exposure to *all* of calculus. However, this separation actually makes calculus more difficult, and prevents students from discovering the connections between differential and integral calculus. We will have no such split in this book, because I believe you can handle the material in one go. Understanding calculus involves figuring out new mathematical concepts like infinity, limits, and summations, but these ideas are not *that* complicated. By getting this far, you've proven you're more than ready to learn the theory, techniques, and applications of derivatives, integrals, sequences, and series.

Let's begin with an overview of the material.

5.2 Overview

This section presents a bird's-eye view of the core concepts of calculus. We'll define more precisely the operations of differentiation and integration, which were introduced in Chapter 2 (see page 133). We'll also discuss the other parts of calculus: *limits*, *sequences*, and *series*. We'll briefly touch upon some applications for each of these concepts; after all, you should know *why* you want to learn all this stuff.

Calculus requires a higher level of abstraction than the mathematical topics discussed in Chapter 1. We began our journey through "Math Land" with the study of *numbers*. Then we learned about *functions*, which are transformations that take real numbers as inputs and produce real numbers as outputs, $f : \mathbb{R} \to \mathbb{R}$. In calculus, the derivative and integral *operators* are procedures that take functions as inputs and produce functions as outputs. Let $\{\mathbb{R} \to \mathbb{R}\}$ denote the set of all functions that take real numbers as inputs and produce real numbers as outputs. The derivative operator takes functions as inputs and produces functions as outputs:

$$\frac{d}{dx} : \{\mathbb{R} \to \mathbb{R}\} \quad \to \quad \{\mathbb{R} \to \mathbb{R}\}.$$

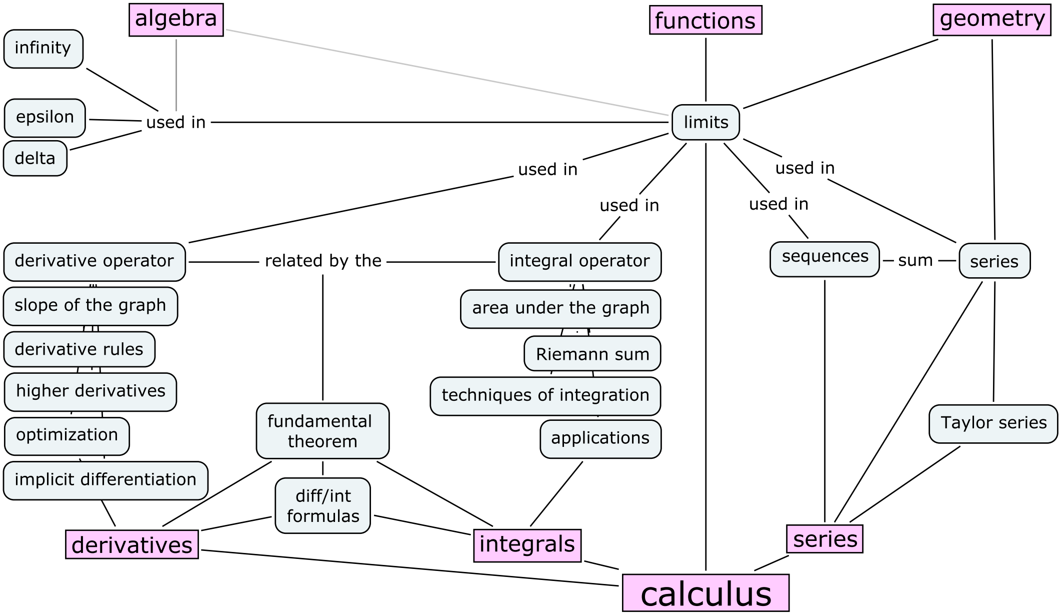


Figure 5.1: The main topics in calculus are limits, derivatives, integrals, sequences, and series. Understanding these notions and how they relate will equip you with many practical problem-solving skills.

More specifically, the derivative operator $\frac{d}{dx}$ acts on a function $f(x)$ to produce its derivative function: $\frac{d}{dx}[f(x)] = f'(x)$.

Differential calculus

Consider the function $f(x)$, which takes real numbers as inputs and produces real numbers as outputs, $f : \mathbb{R} \to \mathbb{R}$. The input variable for the function f is usually denoted x, but we will sometimes also use the variables u, t, and τ to denote the inputs. The function's output is denoted $f(x)$ and is sometimes identified with the y-coordinate in graphs.

The *derivative* function, denoted $f'(x)$ or $\frac{df}{dx}$, describes the *rate of change* of the function $f(x)$. For example, the constant function $f(x) = c$ has derivative $f'(x) = 0$ since the function $f(x)$ does not change at all.

The derivative function describes the *slope* of the graph of the function $f(x)$. The derivative of a line $f(x) = mx + b$ is $f'(x) = m$ since the slope of this line is equal to m. In general, the slope of a function is different at different values of x. For a given choice of input $x = x_0$, the value of the derivative function $f'(x_0)$ is equal to the slope of $f(x)$ as it passes through the point $(x_0, f(x_0))$.

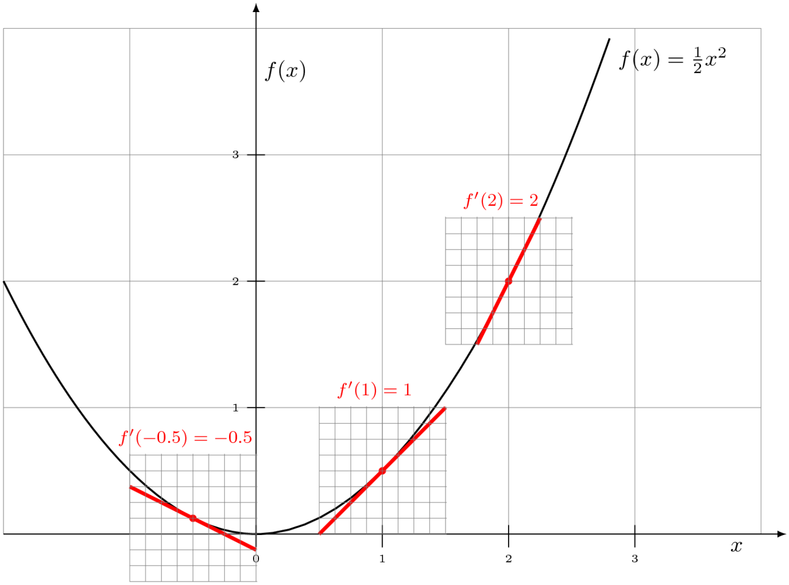


Figure 5.2: The diagram illustrates how to compute the derivative of the function $f(x) = \frac{1}{2}x^2$ at three different points on the graph of the function. To calculate the derivative of $f(x)$ at $x = 1$, we can "zoom in" near the point $(1, 1)$ and draw a line that has the same slope as the function. We can then calculate the slope of the line using a rise-over-run calculation, aided by the mini coordinate system that is provided. The derivative calculations for $x = -0.5$ and $x = 2$ are also shown. Note that the slope of the function is different for each value of x. What is the value of the derivative at $x = 0$? Can you find the general pattern?

The derivative function $f'(x)$ describes the slope of the graph of the function $f(x)$ for all inputs $x \in \mathbb{R}$. The derivative function is a function of the form $f' : \mathbb{R} \to \mathbb{R}$. In our study of mechanics, we learned about the position function $x(t)$ and the velocity function $v(t)$, which describe the motion of an object over time. The velocity is the derivative of the object's position with respect to time $v(t) = \frac{dx}{dt} = x'(t)$.

The derivative function $f'(x)$ is a property of the original function $f(x)$. Indeed, this is where the name *derivative* comes from: $f'(x)$ is not an independent function—it is *derived* from the original function $f(x)$. In mechanics, the function $x(t)$ describes an object's position as a function of time, and the velocity function $v(t)$ describes one property of the position function—namely, how fast the object's position is changing. Similarly, the acceleration function $a(t)$ describes the rate of change of the function $v(t)$.

The *derivative operator*, denoted $\frac{d}{dx}$ or simply D, describes the process of starting from a function $f(x)$ and finding its derivative $f'(x)$. Using the notation for the derivative operator, we can describe the process of "taking the derivative" as follows:

$$f'(x) = \frac{d}{dx} f(x).$$

It is sometimes preferred to describe the derivative in terms of the *derivative operator* $\frac{d}{dx}$, in order to more explicitly show what is going on. There is an original function $f(x)$, and the derivative function $f'(x)$ describes the change in f for a given change in x.

For example, the derivative of the function $f(x) = \frac{1}{2}x^2$ is the function $f'(x) = x$. We can describe this relationship as $(\frac{1}{2}x^2)' = x$ or as $\frac{d}{dx}(\frac{1}{2}x^2) = x$. You should flip back to Figure 5.2 and use the graph to prove to yourself that the slope of $f(x) = \frac{1}{2}x^2$ is described by $f'(x) = x$ everywhere on the graph.

Differentiation techniques

Section 5.6 will formally define the derivative operation. Afterward, we'll develop techniques for computing derivatives, or *taking* derivatives. Computing derivatives is not a complicated task once you learn how to use the derivative formulas. If you flip ahead to Section 5.7 (page 301), you'll find a table of formulas for taking the derivatives of common functions. In Section 5.8, we'll learn the basic rules for computing derivatives of sums, products, and compositions of the basic functions.

Applications of derivatives

Once you develop your ability to find derivatives, you'll be able to use this skill to perform several useful tasks.

Optimization The most prominent application of differential calculus is *optimization*: the process of finding a function's maximum and minimum values. When a function reaches its maximum value, its derivative momentarily becomes zero. The function increases just before it reaches its maximum, and the function decreases just after its maximum. At its maximum value, the function is horizontal, and $f'(x) = 0$ at this point.

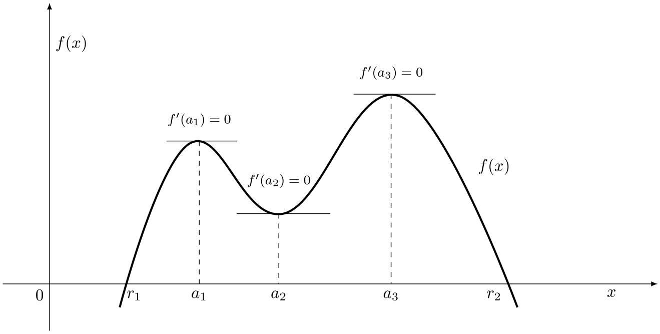


Figure 5.3: The *critical points* of a function occur where the function's derivative equals zero. The critical points of the illustrated function $f(x)$ are $x = a_1$, $x = a_2$, and $x = a_3$. You can use the critical points to find the location of a function's maxima and minima. The point $(a_1, f(a_1))$ is called a *local maximum* of the function, the point at $x = a_2$ is a *local minimum*, while the point at $x = a_3$ is the function's *global maximum*.

The values of x for which $f'(x) = 0$ are called the *critical points* of the function $f(x)$. To find the maximum of a function, we start by compiling a list of its critical points, then go through the list to find the point where the function takes on its largest value. We will discuss the details of this optimization algorithm in Section 5.10.

Tangent lines The *tangent line* to the function $f(x)$ at $x = x_0$ corresponds to the line that passes through the point $(x_0, f(x_0))$ and has the same slope as the function at that point. The word *tangent* comes from the Latin *tangere*, meaning "to touch."

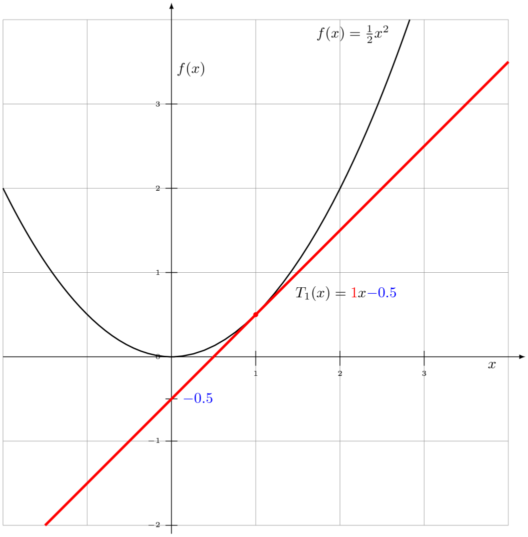


Figure 5.4: An illustration of the tangent line to the function $f(x) = \frac{1}{2}x^2$ at the point $x = 1$. The equation of the tangent line is $T_1(x) = 1x - 0.5$.

The tangent line to the function $f(x)$ at the point $x = x_0$ is described by the equation

$$T_1(x) = \underbrace{f'(x_0)}_{m}\, x \;+\; \underbrace{(f(x_0) - f'(x_0)x_0)}_{b} = f(x_0) \;+\; f'(x_0)(x - x_0).$$

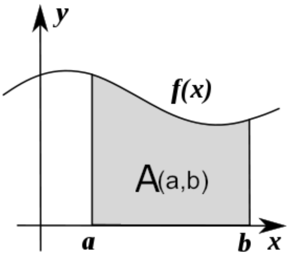
The tangent line $T_1(x)$ is an approximation to the function $f(x)$ near the coordinate $x = x_0$. The approximation $T_1(x)$ is equal to the function $f(x)$ at $x = x_0$ since the tangent line passes through the point $(x_0, f(x_0))$. For coordinates near $x = x_0$, the approximation is also accurate since $T_1(x)$ has the same slope as the function $f(x)$. As the input value x moves farther from x_0, the tangent becomes less accurate at approximating the function $f(x)$.

Integral calculus

The *integral* of $f(x)$ corresponds to the computation of the area under the graph of $f(x)$. The area under $f(x)$ between the points $x = a$ and $x = b$ is denoted as follows:

$$A(a, b) = \int_a^b f(x)\, dx.$$

The area $A(a, b)$ is bounded by the function $f(x)$ from above, by the x-axis from below, and by two vertical lines at $x = a$ and $x = b$. The points $x = a$ and $x = b$ are called the limits of integration. The $\int$ sign comes from the Latin word *summa*. The integral is the "sum" of the values of $f(x)$ between the two limits of integration.



The *integral function* $F(c)$ corresponds to the area calculation as a function of the upper limit of integration:

$$F(c) \equiv \int_0^c f(x)\, dx\,.$$

There are two variables and one constant in this formula. The input variable c describes the upper limit of integration. The *integration variable* x performs a sweep from $x = 0$ until $x = c$. The constant 0 describes the lower limit of integration. Note that choosing $x = 0$ for the starting point of the integral function was an arbitrary choice.

The integral function $F(c)$ contains the "precomputed" information about the area under the graph of $f(x)$. Recall the derivative function $f'(x)$, which tells us the "slope of the graph" property of the function $f(x)$ for all values of x. Similarly, the integral function $F(c)$ tells us the "area under the graph" property of the function $f(x)$ for *all* possible limits of integration.

The area under $f(x)$ between $x = a$ and $x = b$ is obtained by calculating the *change* in the integral function as follows:

$$A(a,b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

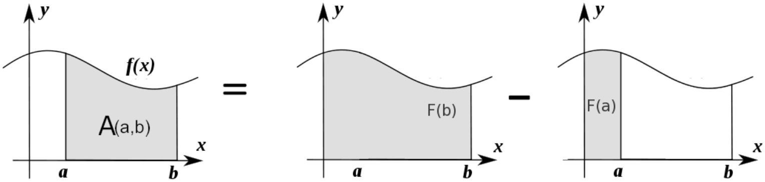


Figure 5.5: The integral function $F(x)$ computes the area under the curve $f(x)$ starting from $x = 0$. The area under $f(x)$ between $x = a$ and $x = b$ is computed using the formula $A(a,b) = F(b) - F(a)$.

Integration techniques

The bulk of the new material needed to understand integral calculus lies in learning various techniques for calculating integrals of functions. Computing integrals is not as easy as computing derivatives, because there are no general rules to follow.

In Section 5.15, we'll describe a number of common techniques for integration. These techniques will enable you to compute the integrals of polynomial functions, exponential functions, logarithmic functions, and trigonometric functions. While these techniques will help you compute integrals in many situations, the process of computing integrals remains somewhat of an art. In art, there are no rules to follow—as an artist, you must be creative and test different approaches until you find one that works.

Applications of integration

Integral calculations have widespread applications to more areas of science than are practical to list here. Let's explore a few examples to gain a general idea of how integrals are applied in the real world.

Computing totals Integral calculations are needed every time we want to compute the total of some quantity that changes over time. If the quantity in question remains constant over time, we can multiply this quantity by the time to find the total quantity. For example, if your monthly rent is $\$720$, your annual rent is $T = \$720 \times 12$.

But what if your rent changes over time? Imagine a crazy landlord who demands you pay on a daily basis and changes the daily rent $r(t)$ each day. Some days rent is $\$20$/day, some days $\$23$/day, and some days he lets you stay for only $\$15$/day. In this situation, computing your annual rent involves the integral $T = \int_0^{365} r(t)\,dt$, which describes the calculation of the daily rate $r(t)$ times the duration of each day dt summed over all the days in the year.

Computing potentials In Section 4.6 we defined the notion of potential energy as the negative of the work done when moving an object against a conservative force. We studied two specific cases: gravitational potential energy $U_g(h) \equiv -\int_0^h \vec{F}_g \cdot d\vec{y} = mgh$, and spring potential energy $U_s(x) \equiv -\int_0^x \vec{F}_s(y) \cdot d\vec{y} = \frac{1}{2}kx^2$. Understanding integrals will allow you to solidify your understanding of the connection between each force $\vec{F}_?(x)$ and its associated potential energy $U_?(x)$.

Computing moments of inertia An object's moment of inertia describes how difficult it is to make the object turn. The moment of inertia is computed as the following integral:

$$I = \int_{\text{obj}} r^2 \, dm.$$

In the mechanics chapter, I asked you to memorize the formulas for $I_{\text{disk}} = \frac{1}{2}mR^2$ and $I_{\text{sphere}} = \frac{2}{5}mR^2$ because it was not yet time to explain the details of integral calculations. After learning about integrals, you'll be able to derive the formulas for I_{disk} and I_{sphere} on your own.

Solving differential equations One of the most important applications of integrals is their ability to "undo" the derivative operation. Recall Newton's second law $F_{\text{net}}(t) = ma(t)$, which can also be written as

$$\frac{F_{\text{net}}(t)}{m} = a(t) = x''(t) = \frac{d}{dx}\left(\frac{d}{dx}x(t)\right).$$

In Chapter 2 we learned how to use integration to solve for $x(t)$ in special cases where the net force is constant $F_{\text{net}}(t) = F_{\text{net}}$. In this chapter, we'll revisit the procedure for finding $x(t)$, and learn how to calculate the motion of an object affected by an external force that varies over time $F_{\text{net}}(t)$.

Limits

The main new tool we'll use in our study of calculus is the notion of a *limit*. In calculus, we often use limits to describe what happens to mathematical expressions when one variable becomes very large, or alternately becomes very small.

For example, to describe a situation where a number n becomes bigger and bigger, we can say,

$$\lim_{n\to\infty} (\text{expression involving } n) .$$

This expression is read, "in the limit as n goes to infinity, expression involving n."

Another type of limit occurs when a small, positive number—for example $\delta > 0$, the Greek letter *delta*—becomes progressively smaller and smaller. The precise mathematical statement that describes what happens when the number δ tends to 0 is

$$\lim_{\delta\to 0} (\text{expression involving } \delta) ,$$

which is read as, "the limit as δ goes to zero, expression involving δ."

Derivative and integral operations are both defined in terms of limits, so understanding limits is essential for calculus. We'll explore limits in more detail and discuss their properties in Section 5.4.

Sequences

So far in this book, we studied functions defined for real-valued inputs $x \in \mathbb{R}$. We can also study functions defined for natural number inputs $n \in \mathbb{N}$. These functions are called *sequences*.

A sequence is a function of the form $a : \mathbb{N} \to \mathbb{R}$. The sequence's input variable is usually denoted n or k, and it corresponds to the *index* or number in the sequence. We describe sequences either by specifying the formula for the n^{th} term in the sequence or by listing all the values of the sequence:

$$a_n, n \in \mathbb{N} \quad \Leftrightarrow \quad (a_0, a_1, a_2, a_3, a_4, \ldots)\,.$$

Note the new notation for the input variable as a subscript. This is the standard notation for describing sequences. Also note the sequence continues indefinitely.

An example of a sequence is

$$a_n = \frac{1}{n^2}\,,\ n \in \mathbb{N}_+ \quad \Leftrightarrow \quad \left(\frac{1}{1}\,,\ \frac{1}{4}\,,\ \frac{1}{9}\,,\ \frac{1}{16}\,,\ \frac{1}{25}\,,\ \ldots\right).$$

This sequence is only defined for strictly positive natural numbers $\mathbb{N}_+ = \{1, 2, 3, 4, \ldots\}$ as the input $n = 0$ yields a divide-by-zero error.

The fundamental question we can ask about sequences is whether they *converge* in the limit when n goes to infinity. For instance, the sequence $a_n = \frac{1}{n^2}$ converges to 0 as n goes to infinity. We can express this fact with the limit expression $\lim\limits_{n\to\infty} \frac{1}{n^2} = 0$.

We'll discuss sequences in more detail in Section 5.18.

Series

Suppose we're given a sequence a_n and we want to compute the sum of all the values in this sequence.

To describe the sum of 3rd, 4th, and 5th elements of the sequence a_n, we turn to summation notation:

$$a_3 + a_4 + a_5 \equiv \sum_{3 \leq n \leq 5} a_n \equiv \sum_{n=3}^{5} a_n \,.$$

The capital Greek letter *sigma* stands in for the word *sum*, and the range of index values included in this sum is denoted below and above the summation sign.

The partial sum of the sequence values a_n ranging from $n = 0$ until $n = N$ is denoted as

$$S_N = \sum_{n=0}^{N} a_n = a_0 + a_1 + a_2 + \cdots + a_{N-1} + a_N.$$

The *series* $\sum a_n$ is the sum of *all* the values in the sequence a_n:

$$\sum a_n \equiv S_\infty = \lim_{N\to\infty} S_N = \sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + a_3 + a_4 + \cdots \,.$$

Note this is an infinite sum.

Techniques

The main mathematical question we'll study with series is the question of their convergence. We say a series $\sum a_n$ *converges* if the infinite sum $S_\infty \equiv \sum_{n\in\mathbb{N}} a_n$ equals some finite number $L \in \mathbb{R}$.

$$S_\infty = \sum_{n=0}^{\infty} a_n = L \quad \Rightarrow \quad \text{the series } \sum a_n \text{ converges.}$$

We call L the *limit* of the series $\sum a_n$.

If the infinite sum $S_\infty \equiv \sum_{n\in\mathbb{N}} a_n$ grows to infinity, we say the series $\sum a_n$ *diverges*.

$$S_\infty = \sum_{n=0}^{\infty} a_n = \pm\infty \quad \Rightarrow \quad \text{the series } \sum a_n \text{ diverges.}$$

The main series technique you need to learn is how to spot the differences between series that converge and series that diverge. You'll soon learn to perform a number of *tests* on the terms in the series, which will indicate whether the infinite sum converges or diverges.

Applications

Series are a powerful computational tool. We can use series to compute approximations to numbers and functions.

For example, the number e can be computed as the following series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2 \cdot 1} + \frac{1}{3 \cdot 2} + \frac{1}{4 \cdot 3 \cdot 2} + \cdots .$$

The factorial operation $n!$ is the product of n times all integers smaller than n: $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$. As we compute more terms from the series, our estimate of the number e becomes more accurate. The partial sum of the first six terms (as shown above) gives us an approximation of e that is accurate to three decimals. The partial sum of the first 12 terms gives us e to an accuracy of nine digits.

Another useful thing you can do with series is approximate functions by infinitely long polynomials. The *Taylor series* approximation for a function $f(x)$ is defined as the series

$$f(x) = \sum_{i=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \cdots .$$

Each term in the series is of the form $a_n = c_n x^n$, where c_n is a constant that depends on the function $f(x)$.

For example, the power series of $\sin(x)$ is

$$\sin(x) = \underbrace{\overbrace{x}^{T_1(x)} - \frac{x^3}{3!} + \frac{x^5}{5!}}_{T_5(x)} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots .$$

Note we can truncate the infinite series anywhere to obtain an approximation to the function. The function $T_5(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$ is the best approximation to the function $\sin(x)$ by a polynomial of degree 5. The equation of the tangent line $T_1(x)$ at $x = 0$ is a special case of the Taylor series approximation procedure, which approximates the function as a first-degree polynomial. We will continue the discussion on series, their properties, and their applications in Section 5.19.

If you haven't noticed yet from glancing at the examples so far, the common theme underpinning all the topics of calculus is the notion of *infinity*. We now turn our attention to the infinite.

5.3 Infinity

Working with infinitely small quantities and infinitely large quantities can be tricky business. It is important that you develop an intuitive understanding of these concepts as soon as possible. Like, now.

Infinitely large

The number ∞ is *really* large. How large? Larger than any number you can think of. Think of any number n. It is true that $n < \infty$. Now think of a bigger number N. It will still hold true that $N < \infty$. In fact, any finite number you can think of, no matter how large, will always be less than ∞.

Technically speaking, ∞ is not a number; infinity is a *process*. You can think of ∞ as the answer you obtain by starting from 0 and continuously adding 1 *forever*.

To see why $N < \infty$ for any finite number N, consider the following reasoning. When we add 1 to a number, we obtain a larger number. The operation $+1$ is equivalent to taking one unit step to the right on the number line. For any n, $n < n + 1$. To get to infinity we start from $n = 0$ and keep adding 1. After N step, we'll arrive at $n = N$. But then we must continue adding 1 and obtain $N + 1$, $N + 2$, $N + 3$, and so on. Since adding 1 always creates a larger number, the following chain of inequalities is true:

$$N \; < \; N + 1 \; < \; N + 2 \; < \; N + 3 < \; \cdots \; < \; \infty.$$

Therefore $N < \infty$ for any finite N.

When we say a number n "goes to" infinity, we are saying n becomes increasingly larger and larger. No number ever actually arrives at infinity since infinity is obtained by adding 1 forever. There is no number $n \in \mathbb{R}$ such that $n = \infty$. Nevertheless, sometimes we can write $N = \infty$, which is an informal way of saying $N = \lim\limits_{n\to\infty} n$.

Infinitely small

The opposite of infinitely large is infinitely small. As a mathematical convention, infinitely small numbers are denoted by the Greek letters ϵ (*epsilon*) and δ (*delta*). The infinitely small number $\epsilon > 0$ is a nonzero number smaller than any number you can think of. The number 0.00001 is pretty small, but it's true that $\epsilon < 0.00001$. The number 10^{-16} extends for 15 zeros after the decimal point, but still ϵ is smaller than it: $\epsilon < 10^{-16}$. Most often, the variable ϵ appears in limit expressions as a quantity that tends toward 0. The expression $\lim\limits_{\epsilon\to 0}$ describes the process of ϵ becoming smaller and smaller, but never actually reaching zero, since by definition $\epsilon > 0$.

Infinitely many

The interval $[0, 1]$ of the number line contains infinitely many numbers. Think of the sequence 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and so forth. There is an infinite number of such fractions, and they all lie in the interval $[0, 1]$.

The ancient Greek philosopher Zeno was confused by this fact. He reasoned as follows. Suppose an archer shoots an arrow and sends it flying toward a target. After some time, the arrow will have travelled half the distance to the target. At some later time, the arrow will have travelled half of the remaining distance and so on, always getting closer to the target. Zeno observed that no matter how little distance remains between the arrow and the target, there will always remain some distance to travel. To reach the target, the arrow would need to pass through an infinite number of points, which is impossible. "How could an infinite number of points fit inside a finite interval?" he figured.

Zeno's argument is not quite right. It is true that the arrow must pass through infinitely many points before it hits the target, but these points "fit" fine in the interval $[0,1]$. These are mathematical points—they don't take up any space at all. We can commend Zeno for thinking about limits centuries before calculus was invented, but we shouldn't repeat his mistake. You must learn how to make limit arguments, because limits are important. Imagine if Zeno had tried to verify his theory experimentally by placing himself in front of an arrow. A wrong argument about limits could get you killed!

Interlude

If the concept of infinity were a person, it would have several problematic character traits. Let's see what we know about infinity so far. The bit about the infinitely large shows signs of megalomania. There is enough of this whole "more, more, more" stuff in the world already, the last thing you want is someone like this as a friend. Conversely, the obsession with the infinitely small ϵ could be a sign of abnormal altruism: the willingness to give up all and leave less and less for oneself. You don't want someone *that* altruistic in your group. And that last part about how infinitely many numbers can fit in a finite interval of the number line sounds infinitely theoretical—definitely not someone to invite to a party.

Let's learn about one redeeming, practical quality of the concept of infinity. Who knows, you might become friends.

Infinitely precise

A computer science (CS) student and a math student are chatting over lunch. The CS student recently learned how to write code that computes mathematical functions as infinitely long series:

$$f(x) = e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2 \cdot 1} + \frac{x^3}{3 \cdot 2} + \frac{x^4}{4 \cdot 3 \cdot 2} + \cdots$$

She wants to tell her friend about her newly acquired powers.

The math student is also learning cool stuff about transcendental numbers. For example the number e can be defined as $e \equiv \lim_{n\to\infty}(1+\frac{1}{n})^n$, but can never be computed exactly—it can only be approximated.

"You know, math is *soooo* much better than CS," says the math student, baiting her friend into an argument about the relative merits of their fields of study.

"What? No way. I can do *anything* on a computer," replies the incredulous scholar of code.

"But can you find exact answers?" the mathematician asks. "Can you compute the number e *exactly*?"

"Sure," says the computer scientist, opening her laptop and typing in a few commands. "The answer is $e = 2.718281828459045$."

"That is not exact," the mathematician points out, "it is just an approximation."

"Whatever—I gave you an approximation to fifteen digits after the decimal. If you're not satisfied with this, then I don't know what your problem is."

"Well, I asked for the *exact* value of e and you only gave me an approximation. Can you find e to 25 digits of precision?" asks the mathematician.

The computer scientist goes back to her laptop.

"Okay, $e = 2.7182818284590452353602875$," she says.

"What about computing e to 50 digits of precision?"

"$e = 2.71828182845904523536028747135266249775724709369995$," says the computer scientist a few seconds later.

"What about—"

"Listen,—" says her friend, "I have this code here that computes e in terms of its power series. The more terms I add in this series, the better my approximation will become. **I can achieve any precision you could ask for**," she explains.

"Then you really know e!" exclaims the mathematician, convinced.

The computer scientist and the mathematician are discussing how to compute approximations to the number e. The mathematician thinks of the number e as the limit $e \equiv \lim\limits_{n\to\infty}(1+\frac{1}{n})^n$. The computer scientist thinks of the number e as the infinite series

$$e = e^1 = \lim_{N\to\infty}\sum_{n=0}^{N}\frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \cdots .$$

Both formulas for e are correct. Observe that we can never compute the value of e exactly, since the formulas for e involve limits to infinity. Because no number ever arrives at infinity, we can never arrive at e either. The number e is a limit. We can only compute numbers that *approach* e.

The computer scientist can obtain approximations to e by computing the partial sum of the first N terms in the series:

$$e_N = \sum_{n=0}^{N}\frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{N!}.$$

Let us denote as ϵ the required precision of the approximation. The more terms she adds, the more accurate the approximation e_N will become. She can always choose a value for N such that the approximation e_N satisfies $|e_N - e| < \epsilon$.

The computer scientist's first answer has a precision of $\epsilon = 10^{-15}$. To obtain an approximation to e with this precision, it is sufficient to compute $N = 19$ terms in the series:

$$e_{19} = \sum_{n=0}^{19}\frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{19!}.$$

The resulting approximation e_{19} is a number somewhere in the interval $(e - 10^{-15}, e + 10^{-15})$. We can also say the absolute value of the difference between e_{19} and the true value of e is smaller than ϵ: $|e_{19} - e| \leq 10^{-15}$.

When the mathematician asks for a precision of $\epsilon' = 10^{-25}$, the computer scientists takes $N = 26$ terms in the series to produce

$$e_{26} = \sum_{n=0}^{26} \frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{19!} + \cdots + \frac{1}{26!},$$

which satisfies $|e_{26} - e| \leq \epsilon'$. In the third step, the mathematician demands a precision $\epsilon'' = 10^{-50}$, and the CS student computes $N = 42$ terms in the series, to produce an approximation satisfying $|e_{42} - e| \leq \epsilon''$. In principle, the game can continue indefinitely because the computer scientist has figured out a *process* for computing increasingly accurate approximations.

This scenario embodies precisely how mathematicians think about limits. It's a bit like a game: the ϵ,N-game. The object of the game is for the CS student to convince the mathematician she knows the number e. The mathematician chooses the precision ϵ. To prove she knows e to precision ϵ, the CS student computes the appropriate number of terms in the series such that her approximation e_N comes ϵ-close to the true answer $|e_N - e| < \epsilon$. If she can produce an approximation which satisfies $|e_N - e| < \epsilon$ **for all** $\epsilon > 0$, then the mathematician will be convinced.

Knowing the value of any finite approximation e_N, no matter how precise, does not constitute a mathematical proof that you can compute e. The mathematician is convinced because the computer scientist has found a *process* for computing approximations with arbitrary precision. In the words of the band Rage Against The Machine,

> "⟨EXPLETITIVE⟩ the G-rides,
> I want the machines that are making them."

Calculus proofs are not about the approximations e_{19}, e_{26}, e_{42}, but about the machines that make them.

The scenarios presented in this section illustrate the need for a precise mathematical language for talking about infinitely large numbers, infinitely small steps, and mathematical procedures with infinite numbers of steps. In the next section we'll learn how to talk about these concepts in terms of *limits*.

5.4 Limits

Limits are the mathematically precise way to talk about infinity. You must understand the language of limits to truly understand the infinitely small, the infinitely large, and the infinitely precise. Once you become comfortable with limits, you'll be able to understand the formal definitions of the derivative and integral operations.

Example

Let's begin with a simple example. Say you have a string of length ℓ and you want to divide it into infinitely many, infinitely short segments. There are infinitely many segments, and they are infinitely short, so together the segments add to the string's total length ℓ.

It's easy enough to describe this process in words. Now let's describe the same process using the notion of a limit. If we divide the length of the string ℓ into N equal pieces then each piece will have a length of

$$\delta \equiv \frac{\ell}{N} .$$

Let's make sure that N pieces of length δ added together equal the string's total length:

$$N\delta = N\frac{\ell}{N} = \ell.$$

Now imagine what happens when the variable N becomes larger and larger. The larger N becomes, the shorter the pieces of string will become. In fact, if N goes to infinity (written $N \to \infty$), then the pieces of string will have zero length:

$$\lim_{N\to\infty} \delta = \lim_{N\to\infty} \frac{\ell}{N} = 0.$$

In the limit as $N \to \infty$, the pieces of string are *infinitely small*.

Note we can still add the pieces of string together to obtain the whole length:

$$\lim_{N\to\infty} (N\delta) = \lim_{N\to\infty} \left(N\frac{\ell}{N} \right) = \ell.$$

Even if the pieces of string are *infinitely small*, because there are *infinitely many* of them, they still add to ℓ.

The notion of a limit is one of the central ideas in this course. As long as you clearly define your limits, you can use infinitely small numbers in your calculations.

Limits at infinity

In math, we're often interested in describing what happens to a certain function when its input variable tends to infinity. This information helps us draw the function's graph. Does $f(x)$ approach a finite number, or does it keep on growing to ∞?

As an example of this type of calculation, consider the limit of the function $f(x) = \frac{1}{x}$ as x goes to infinity:

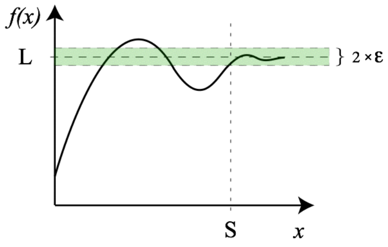
$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \tfrac{1}{x} = 0.$$

This statement is true, even though the function $\frac{1}{x}$ never *actually* reaches zero. The function gets closer and closer to the x-axis but never touches it. This is why the concept of a limit is useful: it allows us to write $\lim\limits_{x\to\infty} f(x) = 0$ even though $f(x) \neq 0$ for any $x \in \mathbb{R}$.

The function $f(x)$ is said to *converge* to the number L if the function approaches the value L for large values of x:

$$\lim_{x\to\infty} f(x) = L.$$

We say "the limit of $f(x)$ as x goes to infinity is the number L." The limit expression is a concise way of saying the following precise mathematical statement: for *any* precision $\epsilon > 0$, there exists a starting point S, after which $f(x)$ equals L within a precision ϵ.



The precise mathematical meaning of $\lim\limits_{x\to\infty} f(x) = L$ is

$$\forall\epsilon > 0\ \exists S \in \mathbb{R} \text{ such that }\ \forall x \geq S \quad |f(x) - L| < \epsilon.$$

I know what you are thinking. Whoa! What just happened here? Chill. I know we saw that upside-down-A and backward-E business all the way back in Chapter 1 (see page 112), so let me rewrite the expression for you in plain English:

For all $\epsilon > 0$,

there exists a number S such that

$|f(x) - L| < \epsilon$ for all x greater than or equal to S.

The limit equation $\lim\limits_{x\to\infty} f(x) = L$ states that the "limit at infinity" of the function $f(x)$ is equal to the number L. This statement is true if and only if there exists a winning strategy for an ϵ,S-game, similar to the ϵ,N-game played by the computer scientist and the mathematician. In the new ϵ,S-game, the mathematician specifies the precision ϵ, and the computer scientists must find a starting point S after which $f(x)$ becomes (and stays) ϵ-close to the limit L. If the computer scientist can succeed for all levels of precision ϵ, then the mathematician will be convinced the equation $\lim\limits_{x\to\infty} f(x) = L$ is true.

Example 2 Calculate $\lim\limits_{x\to\infty} \frac{2x+1}{x}$.

You are given the function $f(x) = \frac{2x+1}{x}$ and must determine what the function looks like for very large values of x. We can rewrite the function as $\frac{2x+1}{x} = 2 + \frac{1}{x}$ to more easily see what is going on:

$$\lim_{x\to\infty} \frac{2x+1}{x} = \lim_{x\to\infty} \left(2 + \frac{1}{x}\right) = 2 + \lim_{x\to\infty} \left(\frac{1}{x}\right) = 2 + 0,$$

since $\frac{1}{x}$ tends toward zero for large values of x.

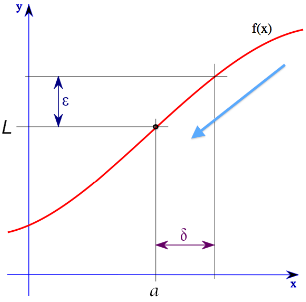
In an introductory calculus course, you will not be required to give formal proofs for statements like $\lim_{x\to\infty} \frac{1}{x} = 0$; instead, you can assume the result is obvious and needs no proof. As the denominator x becomes larger and larger, the fraction $\frac{1}{x}$ becomes smaller and smaller.

Limits to a number

The limit of $f(x)$ approaching $x = a$ *from the right* is defined as

$$\lim_{x \to a^+} f(x) = \lim_{\delta \to 0} f(a + \delta).$$

To find the limit from the right at a, we let x take on values like $a + 0.1$, $a + 0.01$, $a + 0.001$, etc. The diagram shows the graph of a function $f(x)$ near the point $(a, f(a))$. To prove the statement



$$\lim_{x \to a^+} f(x) = L,$$

you must show that

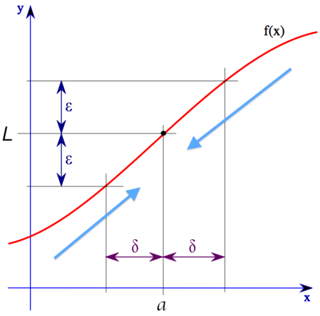
$$\forall \epsilon > 0,\ \exists \delta > 0 \text{ such that } \qquad \forall x \in (a, a + \delta) \quad |f(x) - L| < \epsilon.$$

In other words, the limit from the right corresponds to an ϵ,δ-game in which the mathematician specifies the precision $\epsilon > 0$, and the computer scientist must find a distance $\delta > 0$, such that $|f(x) - L| < \epsilon$, for all x in the range $(a, a + \delta)$.

The limit of $f(x)$ when x approaches *from the left* is defined analogously,

$$\lim_{x \to a^-} f(x) = \lim_{\delta \to 0} f(a - \delta).$$

If both limits from the left and from the right of some number exist and are equal to each other, we can talk about the limit as $x \to a$ without specifying the direction of approach:



$$\lim_{x \to a} f(x) = \lim_{x \to a^+} f(x) = \lim_{x \to a^-} f(x).$$

For the two-sided limit of a function to exist at a point, both limit from the left and the limit from the right

must converge to the same number. If the function $f(x)$
obeys, $f(a) = L$ and $\lim_{x\to a} f(x) = L$, we say the function $f(x)$ is continuous at $x = a$.

Continuity

A function is said to be *continuous* if its graph looks like a smooth curve that doesn't make any sudden jumps and contains no gaps. If you can draw the graph of the function on a piece of paper without lifting your pen, the function is continuous.

A more mathematically precise way to define continuity is to say the function is equal to its limit for all x. We say a function $f(x)$ is *continuous* at a if the limit of f as $x \to a$ converges to $f(a)$:

$$\lim_{x\to a} f(x) = f(a).$$

Remember, the two-sided limit $\lim_{x\to a}$ requires both the left and the right limit to exist and to be equal. Thus, the definition of continuity implies the following equality:

$$\lim_{x\to a^-} f(x) = f(a) = \lim_{x\to a^+} f(x).$$

Consider the mathematical definition of continuity given in the equation above. Can you see how it connects to the intuitive idea of continuous functions as functions that can be drawn without lifting the pen?

Most functions we'll study in calculus are continuous, but not all functions are. Functions that are not defined for some value, as well as functions that make sudden jumps, are not continuous. For example, consider the function

$$f(x) = \frac{|x-3|}{x-3} = \begin{cases} 1 & \text{if } x \geq 3, \\ -1 & \text{if } x < 3. \end{cases}$$

This function is *continuous from the right* at the point $x = 3$, since $\lim_{x\to 3^+} f(x) = 1 = f(3)$. However, taking the limit from the left, we find $\lim_{x\to 3^-} f(x) = -1 \neq f(3)$. Therefore, the function is not continuous. The function $f(x)$ is continuous everywhere on the real line except at $x = 3$.

Example 3 You are asked to calculate $\lim\limits_{x\to 5} \dfrac{2x+1}{x}$.

$$\lim_{x\to 5} \frac{2x+1}{x} = \frac{2(5)+1}{5} = \frac{11}{5}.$$

There is nothing tricky going on here—plug the number 5 into the equation, and voila. The function $f(x) = \frac{2x+1}{x}$ is continuous at the value $x = 5$, so the limit of the function as $x \to 5$ is equal to the value of the function $\lim\limits_{x\to 5} f(x) = f(5)$.

Asymptotes

An *asymptote* of the function $f(x)$ is a line the function approaches but never touches. The word asymptote comes from the Greek *asumptotos*, which means "not falling together." For example, the line $y = 0$ (the x-axis) is an asymptote of the function $f(x) = \frac{1}{x}$ as x goes to infinity.

A *vertical asymptote* is a vertical line that the function approaches. For example, the function $f(x) = \frac{1}{3-x}$ has a vertical asymptote at $x = 3$. When the function approaches $x = 3$ from the left, the function increases to infinity:

$$\lim_{x\to 3^-} \frac{1}{3-x} = \infty.$$

The limit describes x taking on values like 2.9, 2.99, 2.999, and so on. The number in the denominator gets smaller and smaller, thus the fraction grows larger and larger. Note, the function is not defined at the exact value $x = 3$. Nevertheless, the above limit allows us to describe what happens to the function near that point.

Example 4 Find $\lim_{x\to 0} \frac{2x+1}{x}$.

Plugging $x = 0$ into the fraction yields a divide-by-zero error $\frac{2(0)+1}{0}$, so a more careful treatment is required.

First we'll consider the limit from the right $\lim_{x\to 0+} \frac{2x+1}{x}$. We want to approach the value $x = 0$ with small positive numbers. First we'll define a small positive number $\delta > 0$, then choose $x = \delta$, and then compute the limit:

$$\lim_{\delta \to 0} \frac{2(\delta)+1}{\delta} = 2 + \lim_{\delta \to 0} \frac{1}{\delta} = 2 + \infty = \infty.$$

In this instance, we take it for granted that $\lim_{\delta \to 0} \frac{1}{\delta} = \infty$. Intuitively, let's imagine what happens in the limit as δ approaches 0. When $\delta = 10^{-3}$, the function value will be $\frac{1}{\delta} = 10^3$. When $\delta = 10^{-6}$, $\frac{1}{\delta} = 10^6$. As $\delta \to 0$, the expression $\frac{1}{\delta}$ becomes larger and larger all the way to infinity.

If we take the limit from the left, letting x take on small negative values, we obtain

$$\lim_{\delta \to 0} f(-\delta) = \frac{2(-\delta)+1}{-\delta} = -\infty.$$

Since $\lim_{x \to 0^+} f(x)$ does not equal $\lim_{x \to 0^-} f(x)$, we say $\lim_{x \to 0} f(x)$ does not exist.

Limits are fundamentally important for calculus. Indeed, the three main calculus topics we'll discuss in the remainder of this chapter are derivatives, integrals, and series—all of which are defined using limits.

Limits for derivatives

The formal definition of a function's derivative is expressed in terms of the rise-over-run formula for an infinitesimally short run:

$$f'(x) \ = \ \lim_{\text{run} \to 0} \frac{\textsf{rise}}{\textsf{run}} \ = \ \lim_{\delta \to 0} \frac{f(x+\delta) \ - \ f(x)}{x+\delta \ - \ x}.$$

We'll continue the discussion of this formula in Section 5.6.

Limit for integrals

One way to approximate the area under the curve $f(x)$ between $x = a$ and $x = b$ is to split the area into N little rectangles of width $\epsilon = \frac{b-a}{N}$ and height $f(x)$, and then calculate the sum of the areas of the rectangles:

$$A(a,b) \approx \underbrace{\epsilon f(a) + \epsilon f(a+\epsilon) + \epsilon f(a+2\epsilon) + \cdots + \epsilon f(b-\epsilon)}_{N \text{ terms}}.$$

We obtain the exact value of the area in the limit where we split the area into an infinite number of rectangles with infinitely small width:

$$\int_a^b f(x)\,dx = A(a,b) = \lim_{N\to\infty} \left[\epsilon f(a) + \epsilon f(a+\epsilon) + \epsilon f(a+2\epsilon) + \cdots + \epsilon f(b-\epsilon)\right].$$

Computing the area under a function by splitting the area into infinitely many rectangles is an approach known as a *Riemann sum*, which we'll discuss in Section 5.13.

Limits for series

We use limits to describe the convergence properties of series. For example, the partial sum of the first N terms of the geometric series $a_n = r^n$ corresponds to the following expression:

$$S_N = \sum_{n=0}^{N} r^n = 1 + r + r^2 + r^3 + \cdots + r^N.$$

The *series* a_n is defined as the limit $N \to \infty$ of the above expression. For values of r that obey $|r| < 1$, the series converges:

$$S_\infty = \lim_{N\to\infty} S_N = \sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}.$$

To convince yourself the above formula is correct, observe how the infinite sum S_∞ is similar to a shifted version of itself: $S_\infty = 1 + rS_\infty$. Now solve for S_∞ in the equation $S_\infty = 1 + rS_\infty$.

You'll find more about series in Section 5.19.

5.5 Limit formulas

We now switch gears into *reference* mode, where we'll describe a whole bunch of known formulas for limits of various kinds of functions. You do not need to know *why* these limit formulas are true; your mission is to understand what they mean.

Ratios of functions

The following statements tell you about the *relative sizes* of functions. If the limit of the ratio of two functions is equal to 1, then these functions must behave similarly in the limit. If the limit of the ratio goes to zero, then the function in the denominator must be much larger than the function in the numerator.

Limits of trigonometric functions:

$$\lim_{x\to 0}\frac{\sin(x)}{x}=1,\quad \lim_{x\to 0}\cos(x)=1,\quad \lim_{x\to 0}\frac{1-\cos x}{x}=0,\quad \lim_{x\to 0}\frac{\tan(x)}{x}=1.$$

A polynomial of degree n and the exponential function base a with $a>1$ both go to infinity as x goes to infinity:

$$\lim_{x\to\infty} x^n=\infty,\qquad\qquad \lim_{x\to\infty} a^x=\infty.$$

Though both functions grow to infinity, the exponential function grows much faster. The limit of the ratio of the exponential function divided by any polynomial function is

$$\lim_{x\to\infty}\frac{a^x}{x^n}=\infty,\qquad \text{for all } n\in\mathbb{N}, |a|>1.$$

In computer science, this distinction is a big deal when comparing the running times of algorithms. Imagine x represents the *size* of the problem we want to solve. A *polynomial-time algorithm* will take fewer than Cx^n steps to compute the answer, for some constants C and n. An *exponential-time algorithm* takes an exponential number of steps to compute the answer—the number of steps is described by the expression Da^x, for some constants D and a. Exponential-time algorithms are kind of useless because their running time becomes prohibitively long for large problems.

With a large enough input x, an exponential-time algorithm with running time Da^x will take longer than the age of the universe to finish!

Euler's number

The number e is defined as the following limit:

$$e \equiv \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \quad \text{or alternately as} \quad e = \lim_{\epsilon\to 0}(1+\epsilon)^{1/\epsilon}.$$

The first expression corresponds to a compound interest calculation with an annual interest rate of 100% where compounding is performed infinitely often.

For future reference, here are some other limits involving the exponential function:

$$\lim_{x\to 0}\frac{\mathrm{e}^x-1}{x}=1, \qquad \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n=\mathrm{e}^x.$$

And here are some limits involving logarithms:

$$\lim_{x\to 0^+} x^a\ln(x)=0, \qquad \lim_{x\to\infty}\frac{\ln^p(x)}{x^a}=0,\ \forall p<\infty$$

$$\lim_{x\to 0}\frac{\ln(x+a)}{x}=a, \qquad \lim_{x\to 0}\left(a^{1/x}-1\right)=\ln(a).$$

Properties

The calculation of the limit of the sum, difference, product, and quotient of two functions is computed as follows, respectively:

$$\begin{aligned}
\lim_{x\to a}(f(x)+g(x)) &= \lim_{x\to a}f(x)+\lim_{x\to a}g(x),\\
\lim_{x\to a}(f(x)-g(x)) &= \lim_{x\to a}f(x)-\lim_{x\to a}g(x),\\
\lim_{x\to a}f(x)g(x) &= \lim_{x\to a}f(x)\cdot\lim_{x\to a}g(x),\\
\lim_{x\to a}\frac{f(x)}{g(x)} &= \frac{\lim\limits_{x\to a}f(x)}{\lim\limits_{x\to a}g(x)}.
\end{aligned}$$

The above formulas indicate we are allowed to *take the limit inside* of the basic arithmetic operations.

L'Hopital's rule

If you are taking the limit of a fraction of two functions $\frac{f(x)}{g(x)}$ that obey $\lim\limits_{x\to\infty} f(x) = 0$ and $\lim\limits_{x\to\infty} g(x) = \infty$, then the limit of their ratio is

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to\infty} f(x)}{\lim\limits_{x\to\infty} g(x)} = \frac{0}{\infty} = 0.$$

Both the numerator and the denominator help drive the ratio to zero. Alternately, if you ever obtain a fraction of the form $\frac{\infty}{0}$ as a limit, where both the large numerator and the small denominator make the fraction grow to infinity, you can write $\frac{\infty}{0} = \infty$.

Sometimes, when evaluating limits of fractions $\frac{f(x)}{g(x)}$, you might end up with a fraction like

$$\frac{0}{0} \qquad \text{or} \qquad \frac{\infty}{\infty}.$$

These are called *undecidable* conditions. They are undecidable because we cannot tell whether the function in the numerator or the denominator is bigger. One way to compute limits with undecidable conditions is to compare the ratio of the derivatives of the numerator and the denominator. This is called L'Hopital's rule:

$$\lim_{x\to a} \frac{f(x)}{g(x)} \quad \overset{\text{H.R.}}{=} \quad \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

You can find the derivative formulas you'll need for using L'Hopital's rule in the table of derivative formulas on page 301.

Example Consider the calculation of the limit of the ratio $\frac{x^3}{e^x}$ as x goes to infinity. Both functions grow to infinity. We can calculate the limit of their ratio by using

L'Hopital's rule three times:

$$\lim_{x\to\infty} \frac{x^3}{e^x} \overset{\text{H.R.}}{=} \lim_{x\to\infty} \frac{3x^2}{e^x} \overset{\text{H.R.}}{=} \lim_{x\to\infty} \frac{6x}{e^x} \overset{\text{H.R.}}{=} \lim_{x\to\infty} \frac{6}{e^x} = \frac{6}{\infty} = 0.$$

Example 2 Calculate the limit $\lim_{x\to 0} \frac{\sin^{-1}(x)}{x}$. Both the numerator and the denominator go to zero as x goes to zero. We can find the derivative formula for $\sin^{-1}(x)$ in the table on page 301, then apply L'Hopital's rule:

$$\lim_{x\to 0} \frac{\sin^{-1}(x)}{x} \overset{\text{H.R.}}{=} \lim_{x\to 0} \frac{\frac{1}{\sqrt{1-x^2}}}{1} = \lim_{x\to 0} \frac{1}{\sqrt{1-x^2}} = \frac{1}{\sqrt{1-0}} = 1.$$

Links

[See the Wikipedia page for more examples of limits]
`https://en.wikipedia.org/wiki/Limit_of_a_function`

5.6 Derivatives

In the beginning of the chapter we introduced the derivative concept by identifying the derivative with the slope of the function's graph. This graphical representation of derivatives and the intuition that comes with it are very important: this is how mathematicians and physicists usually "think" about derivatives. It is equally important to understand the formal definition of the derivative operation, so this is what we will cover next. After, we will build some practical skills for calculating derivatives of functions.

Definition

The derivative of a function is defined as

$$f'(x) \equiv \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}.$$

The definition of the derivative comes from the rise-over-run formula for calculating the slope of a line:

$$\frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x} = \frac{y_f - y_i}{x_f - x_i} = \frac{f(x+\delta) \ - \ f(x)}{x + \delta \ - \ x}.$$

By making δ tend to zero in the above expression, we are able to obtain the slope of the function $f(x)$ at the point x.

Derivatives occur so often in math that people have devised many ways to denote them. Don't be fooled by this multitude of notations—all of them refer to the same concept:

$$Df(x) \equiv f'(x) \equiv \frac{d}{dx}f(x) \equiv \frac{df}{dx} \equiv \dot{f} \equiv \nabla f.$$

Example Let's calculate the derivative of $f(x) = 2x^2 + 3$ to illustrate how the complicated-looking derivative formula works:

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta} = \lim_{\delta \to 0} \frac{2(x+\delta)^2 + 3 \quad - \quad (2x^2 + 3)}{\delta}.$$

We can simplify the fraction inside the limit:

$$\frac{2x^2 + 4x\delta + \delta^2 - 2x^2}{\delta} = \frac{4x\delta + \delta^2}{\delta} = \frac{4x\delta}{\delta} + \frac{\delta^2}{\delta} = 4x + \delta.$$

The second term of this expression disappears when we take the limit to obtain the final answer:

$$f'(x) = \lim_{\delta \to 0} (4x + \delta) = 4x + 0 = 4x.$$

Congratulations, you have just calculated your first derivative! The calculation wasn't that complicated, but the process was pretty long and tedious. The good news is you only need to calculate the derivative from first principles once. Once you obtain a *derivative formula* for a particular function, you can use the formula every time you see a function of that form.

The power rule

The derivative formula we obtained in the last example is a special case of the general formula for computing derivatives of powers of x. The *power rule* formula states:

$$\text{if} \quad f(x) = x^n \quad \text{then} \quad f'(x) = nx^{n-1}.$$

The proof of this formula proceeds by steps analogous to the steps used in the example above.

Example 2 Use the power rule to compute the derivatives of the following functions:

$$f(x) = x^{10}, \qquad g(x) = \sqrt{x^3}, \qquad h(x) = \frac{1}{x^3}\,.$$

In the first case, we apply the formula directly to find the derivative $f'(x) = 10x^9$. In the second case, we begin with the fact that square root is equivalent to an exponent of $\frac{1}{2}$, thus we rewrite the function as $g(x) = x^{\frac{3}{2}}$. After rewriting, we find $g'(x) = \frac{3}{2}x^{\frac{1}{2}} = \frac{3}{2}\sqrt{x}$. We can rewrite the third function as $h(x) = x^{-3}$, then use the power rule to compute the derivative $h'(x) = -3x^{-4} = -\frac{3}{x^4}$.

Applications of derivatives

Optimization

Consider some real-world problem in which a quantity is described by the function $f(x)$. The derivative function $f'(x)$ describes how the quantity *changes* over time. Often, we don't actually care about the value of $f'(x)$ and only need to find the sign of the derivative. If the derivative is positive $f'(x) > 0$, the function is *increasing*. If $f'(x) < 0$, the function is *decreasing*. If the function is horizontal at a certain point $x = x_0$, then $f'(x_0) = 0$. The points where $f'(x) = 0$ are important for finding the maximum and minimum values of $f(x)$.

Recall we previously used the rule "the max is where the derivative is zero" to calculate the maximum height h reached by a ball thrown in the air. We identified the top of the ball's trajectory as the location when its velocity in the y-direction equals zero. The ball moves upward initially ($v_y > 0$), stops momentarily at its maximum height ($v_y = 0$), then moves downward ($v_y < 0$) until it comes back to the ground. We can find the time t_{top} it takes for the ball to reach its top height by solving the equation $v_y(t_{\text{top}}) = 0$, which is the same as $y'(t_{\text{top}}) = 0$. Once we know the time t_{top}, we substitute this value into the equation for $y(t)$ to obtain $h = \max\{y(t)\} = y(t_{\text{top}})$.

We'll discuss the details of the *optimization algorithm* in Section 5.10.

Tangent lines

The *tangent line* to the function $f(x)$ at $x = x_0$ is the line with slope $f'(x_0)$ that passes through the point $(x_0, f(x_0))$. The tangent line is special because it only "touches" the function at a single point. This is in contrast with *secant* lines (from the Latin *secare*), which *cut* through the function at more than one point. There are infinitely many secant lines that cut through the point $(x_0, f(x_0))$, but only a single tangent line—denoted $T_1(x)$.

Let's calculate the equation of the tangent line at $x = x_0$. We are looking for the equation of the line $T_1(x) = mx + b$ that passes through the point $(x_0, f(x_0))$ and has slope equal to $f'(x_0)$. Since m represents the slope of the line,

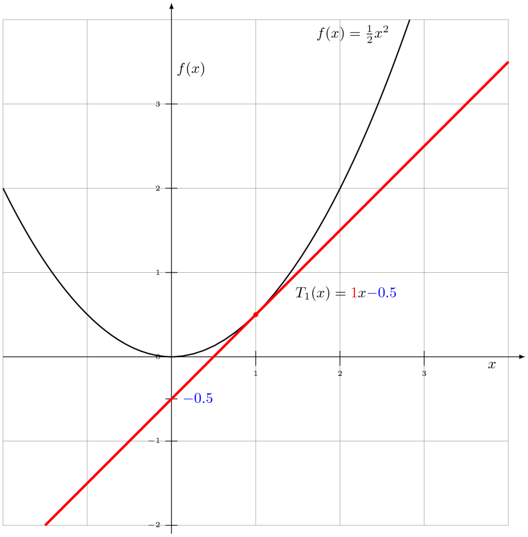


Figure 5.6: The tangent line to the function $f(x) = \frac{1}{2}x^2$ at the point $x = 1$ is $T_1(x) = 1x - 0.5$. Note how the tangent line touches the function *only* at the point $(1, 1)$ and the slope of the tangent line is the same as the slope of $f(x)$ at that point.

we can conclude that $m = f'(x_0)$. Since the tangent line passes through the point $(x_0, f(x_0))$, we can find the initial value b by solving the equation

$$f(x_0) = T_1(x_0) = f'(x_0)x_0 + b.$$

Solving for b, we find $b = f(x_0) - f'(x_0)x_0$.

The equation of a tangent line can be written as

$$T_1(x) = \underbrace{f(x_0)}_{\text{(c)}} + \underbrace{f'(x_0)(x - x_0)}_{(\ell)}.$$

The above expression describes the tangent line as the sum of a constant term (c) and a second term (ℓ) proportional to the shifted coordinate $(x - x_0)^1$ centred at x_0.

The tangent line $T_1(x)$ is the best linear approximation to the function $f(x)$ near the coordinate $x = x_0$. Written informally, this statement says,

$$f(x) \approx T_1(x) \quad \text{for } x \text{ near } x_0.$$

We previously used this type of linear approximation to derive simple harmonic motion equations for a pendulum on page 254. The *small angle* approximation states that

$$f(\theta) = \sin\theta \approx \theta = T_1(\theta), \qquad \text{for } \theta \text{ near } 0.$$

Discussion

Now that you know what derivatives are and what they are used for, it's time to learn how to compute them.

5.7 Derivative formulas

The table below shows the derivative formulas for a number of commonly used functions. You'll be using these derivative formulas a lot in the remainder of this chapter so it's a good idea to memorize them.

$f(x)$	$-$ derivative $\rightarrow$	$f'(x)$
a		0
$\alpha f(x) + \beta g(x)$		$\alpha f'(x) + \beta g'(x)$
x		1
x^n		nx^{n-1}
$\frac{1}{x} \equiv x^{-1}$		$\frac{-1}{x^2} \equiv -x^{-2}$
$\sqrt{x} \equiv x^{\frac{1}{2}}$		$\frac{1}{2\sqrt{x}} \equiv \frac{1}{2}x^{-\frac{1}{2}}$
e^x		e^x

a^x	$a^x \ln(a)$
$\ln(x)$	$\dfrac{1}{x}$
$\log_a(x)$	$(x \ln(a))^{-1}$
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\tan(x)$	$\sec^2(x) \equiv \cos^{-2}(x)$
$\sin^{-1}(x)$	$\dfrac{1}{\sqrt{1-x^2}}$
$\cos^{-1}(x)$	$\dfrac{-1}{\sqrt{1-x^2}}$
$\tan^{-1}(x)$	$\dfrac{1}{1+x^2}$
$\sinh(x)$	$\cosh(x)$
$\cosh(x)$	$\sinh(x)$

You can find a complete table of derivative formulas on page 462 in the back of the book.

5.8 Derivative rules

Taking derivatives is a simple task: find the appropriate formula in the table of derivative formulas and apply the formula to the specific problem at hand. Derivative tables come in handy, but they usually do not list formulas for *composite* functions. This section covers some important derivatives rules that will allow you to find derivatives of more complicated functions.

Linearity

The derivative of a sum of two functions is the sum of the derivatives:

$$[f(x)+g(x)]' = f'(x)+g'(x),$$

and for any constant α, we have

$$[\alpha f(x)]' = \alpha f'(x).$$

The derivative of a linear combination of functions $\alpha f(x) + \beta g(x)$ is equal to the linear combination of the derivatives $\alpha f'(x) + \beta g'(x)$.

Product rule

The derivative of a product of two functions is obtained as follows:

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x).$$

Quotient rule

The quotient rule tells us how to obtain the derivative of a fraction of two functions:

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

Chain rule

If you encounter a situation that includes an inner function and an outer function, like $f(g(x))$, you can obtain the derivative by a two-step process:

$$[f(g(x))]' = f'(g(x))g'(x).$$

In the first step, leave the inner function $g(x)$ alone. Focus on taking the derivative of the outer function $f(x)$, and leave the expression $g(x)$ inside the f' expression. As the second step, multiply the resulting expression by the derivative of the *inner* function $g'(x)$.

The chain rule tells us the derivative of a composite function is calculated as the product of the derivative of the outer function and the derivative of the inner function.

Example

$$\left[\sin(x^2)\right]' = \cos(x^2)\left[x^2\right]' = \cos(x^2)2x.$$

More complicated example The chain rule also applies to functions of functions of functions $f(g(h(x)))$. To take the derivative, start from the outermost function and work your way toward x.

$$[f(g(h(x)))]' = f'(g(h(x)))g'(h(x))h'(x).$$

Now let's try taking another derivative

$$\left[\sin(\ln(x^3))\right]'=\cos(\ln(x^3))\left[\ln(x^3)\right]'=\cos(\ln(x^3))\frac{1}{x^3}\left[x^3\right]'=\cos(\ln(x^3))\frac{3}{x}.$$

Simple, right?

Examples

The above rules define *all* you need to know to take the derivative of any function, no matter how complicated. To convince you, I'll show you some examples of really hairy functions. Don't be scared by complexity: as long as you follow the rules, you'll find the right answer in the end.

Example Calculate the derivative of

$$f(x) = e^{x^2}.$$

We need the chain rule for this one:

$$f'(x) = e^{x^2}[x^2]' = e^{x^2}2x.$$

Example 2 Find the derivative of

$$f(x) = \sin(x)e^{x^2}.$$

We'll need the product rule for this one:

$$f'(x) = \cos(x)e^{x^2} + \sin(x)2xe^{x^2}.$$

Example 3 Compute the derivative of

$$f(x) = \sin(x)e^{x^2}\ln(x).$$

This situation again calls for the product rule, but this time we'll have three terms. For each term, take the derivative of one of the functions and multiply this derivative by the other two functions:

$$f'(x) = \cos(x)e^{x^2}\ln(x) + \sin(x)2xe^{x^2}\ln(x) + \sin(x)e^{x^2}\frac{1}{x}.$$

Example 4 Take the derivative of

$$f(x) = \sin(\cos(\tan(x)))\,.$$

We need a triple chain rule for this one:

$$\begin{aligned} f'(x) &= \cos(\cos(\tan(x)))\,[\cos(\tan(x))]' \\ &= -\cos(\cos(\tan(x)))\sin(\tan(x))\,[\tan(x)]' \\ &= -\cos(\cos(\tan(x)))\sin(\tan(x))\sec^2(x). \end{aligned}$$

Explanations

Derivation of the product rule

By definition, the derivative of $f(x)g(x)$ is

$$[f(x)g(x)]' = \lim_{\delta\to 0}\frac{f(x+\delta)g(x+\delta) - f(x)g(x)}{\delta}.$$

Consider the numerator of the fraction. If we add and subtract $f(x)g(x+\delta)$, we can factor the expression into two terms, like this:

$$\begin{gathered} f(x+\delta)g(x+\delta)\ \overbrace{-f(x)g(x+\delta) + f(x)g(x+\delta)}^{=0}\ -f(x)g(x) \\ = [f(x+\delta) - f(x)]g(x+\delta) + f(x)[g(x+\delta) - g(x)]. \end{gathered}$$

The expression for the derivative of the product becomes

$$[f(x)g(x)]' = \left\{ \lim_{\delta\to 0} \frac{[f(x+\delta)-f(x)]}{\delta} g(x+\delta) + f(x)\frac{[g(x+\delta)-g(x)]}{\delta} \right\}.$$

This looks almost exactly like the product rule formula, except here we have $g(x+\delta)$ instead of $g(x)$. This difference is okay since we assume $g(x)$ is a continuous function. Recall that a continuous function $g(x)$ obeys $\lim_{\delta\to 0} g(x+\delta) = g(x)$ for all x. Using the continuity property of $g(x)$, we obtain the final form of the product rule:

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x).$$

Derivation of the chain rule

Before we begin, I'd like to remark on the notation used to define derivatives. I happen to like the Greek letter δ (*delta*), so I defined the derivative of $f(x)$ as

$$f'(x) = \lim_{\delta\to 0} \frac{f(x+\delta)-f(x)}{\delta}.$$

Instead, I could have used the variable h and written

$$f'(x) \equiv \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}.$$

In fact, we can use *any* variable. All that matters is that we divide by the *same* quantity as the quantity added to x inside the function, and that this quantity goes to zero.

The derivative of $f(g(x))$ is

$$[f(g(x))]' = \lim_{\delta\to 0} \frac{f(g(x+\delta))-f(g(x))}{\delta}.$$

The trick is to define a new quantity

$$\Delta = g(x+\delta) - g(x),$$

and then substitute $g(x+\delta) = g(x) + \Delta$ into the derivative expression:

$$[f(g(x))]' = \lim_{\delta\to 0} \frac{f(g(x)+\Delta) - f(g(x))}{\delta}.$$

This is starting to look more like a derivative formula, but the quantity added in the input is different from the quantity by which we divide. To fix this, we can multiply and divide by Δ and rearrange the expression to obtain

$$\lim_{\delta\to 0} \frac{f(g(x)+\Delta) - f(g(x))}{\delta}\frac{\Delta}{\Delta} = \lim_{\delta\to 0} \frac{f(g(x)+\Delta) - f(g(x))}{\Delta}\frac{\Delta}{\delta}.$$

Now use the definition of the quantity Δ and rearrange the fraction:

$$[f(g(x))]' = \lim_{\delta\to 0} \frac{f(g(x)+\Delta) - f(g(x))}{\Delta}\frac{g(x+\delta) - g(x)}{\delta}.$$

This looks a lot like $f'(g(x))g'(x)$, and in fact, it is. Taking the limit $\delta \to 0$ implies that the quantity $\Delta(\delta) \to 0$. This is because the function $g(x)$ is continuous: $\lim_{\delta\to 0}[g(x+\delta) - g(x)] = 0$. Taking a derivative by using the quantity Δ is just as good as using δ. Thus, we've shown that

$$[f(g(x))]' = f'(g(x))g'(x).$$

Alternate notation

The presence of so many primes and brackets can make the expressions above difficult to read. As an alternative, we sometimes use another notation for derivatives. The three rules of derivatives in the alternate notation are written as follows:

- Linearity: $\frac{d}{dx}(\alpha f(x) + \beta g(x)) = \alpha\frac{df}{dx} + \beta\frac{dg}{dx}$
- Product rule: $\frac{d}{dx}(f(x)g(x)) = \frac{df}{dx}g(x) + f(x)\frac{dg}{dx}$
- Chain rule: $\frac{d}{dx}\left(f(g(x))\right) = \frac{df}{dg}\frac{dg}{dx}$

Some authors prefer the notation $\frac{df}{dx}$ for the derivative of the function $f(x)$, because it more evocative of a rise-over-run calculation.

5.9 Higher derivatives

In the previous section we learned how to calculate the derivative $f'(x)$ of any function $f(x)$. The second derivative of $f(x)$ is the derivative of the derivative of $f(x)$, and is denoted as

$$f''(x) \equiv [f'(x)]' \equiv \frac{d}{dx}f'(x) \equiv \frac{d^2}{dx^2}f(x).$$

This process can be continued to calculate higher derivatives of $f(x)$.

In practice, the first and second derivatives are most important because they have a geometrical interpretation. The first derivative of $f(x)$ describes the *slope* of $f(x)$ while the second derivative describes the *curvature* of $f(x)$.

Definitions

- $f(x)$: the original function
- $f'(x)$: the first derivative of the function $f(x)$. The first derivative contains information about the *slope* of the function $f(x)$.
- $f''(x)$: the second derivative of the function $f(x)$. The second derivative contains information about the *curvature* of the function $f(x)$.
  - ▷ If $f''(x) > 0$ for all x, the function $f(x)$ is *convex*.
    Convex functions open upward, like $f(x) = x^2$.
  - ▷ If $f''(x) < 0$ for all x, the function $f(x)$ is *concave*.
    Concave functions open downward, like $f(x) = -x^2$.
- $f'''(x) \equiv f^{(3)}(x)$: the third derivative of $f(x)$
- $f^{(n)}(x)$: the nth derivative of $f(x)$

Second derivative

The second derivative describes the change in the value of the first derivative. To obtain $f''(x)$ we compute the derivative of $f'(x)$.

The second derivative tells us about the *curvature* of the function $f(x)$. If the curvature of a function is positive ($f''(x) > 0$), this means the function's slope is increasing, so the function must curve upward. Negative curvature means the function curves downward.

Example Calculate the second derivatives of the functions $u(x) = x^2$ and $d(x) = -x^2$ and comment on the *shape* of these functions.

To solve this problem, we calculate the first derivatives $u'(x) = 2x$ and $d'(x) = -2x$. We obtain a function's second derivative by taking the derivative of its first derivative: $u''(x) = 2$ and $d''(x) = -2$. Since the second derivative of the function $u(x)$ is always positive, the curvature of the function $u(x)$ is always positive. We say the function $u(x)$ is *convex*; it opens upward. On the other hand $d(x)$ is *concave*; it opens downward.

The functions $u(x)$ and $d(x)$ are canonical examples of functions with positive and negative curvature. If a function $f(x)$ has positive curvature at a point x^* ($f''(x^*) > 0$), then the function locally *resembles* $u(x - x^*) = (x - x^*)^2$. On the other hand, if the second derivative of $f(x)$ is negative at x^*, the function will locally resemble $d(x - x^*) = -(x - x^*)^2$. In other words, the terms *convex* and *concave* refer to the u-likeness vs. d-likeness property of functions.

Higher derivatives

If we take the derivative of the derivative of the derivative of $f(x)$, we obtain the *third* derivative of the function. This process can be continued further to obtain the n$^{\text{th}}$ derivative of the function:

$$f^{(n)}(x) \equiv \frac{d^n}{dx^n} f(x) \equiv \underbrace{\frac{d}{dx}\frac{d}{dx}\cdots\frac{d}{dx}}_{n} f(x).$$

Higher derivatives do not have an obvious geometrical interpretation. However, if you are given a function $f(x)$ such that $f'''(x) > 0$, then the function $f(x)$ must be $+x^3$-like. Alternately, if $f'''(x) < 0$, then the function must resemble $-x^3$.

Later in this chapter, we will learn how to compute the Taylor series of a function, which is a procedure used to find polynomial approximations to any function $f(x)$:

$$f(x) \ \approx \ c_0 + c_1x + c_2x^2 + c_3x^3 + c_4x^4 + \cdots + c_nx^n.$$

The values of the coefficients c_0, c_1, ..., c_n in the approximation require us to compute higher derivatives of $f(x)$. The coefficient c_n tells us whether $f(x)$ is more similar to $+x^n$ ($c_n > 0$), or to $-x^n$ ($c_n < 0$), or to neither of the two ($c_n = 0$).

Example Compute the third derivative of $f(x) = \sin(x)$.

The first derivative is $f'(x) = \cos(x)$. The second derivative will be $f''(x) = -\sin(x)$ so the third derivative must be $f'''(x) = -\cos(x)$. Note that $f^{(4)}(x) = f(x)$.

Optimization: the killer app of calculus

Knowing your derivatives will allow you to *optimize* any function—a crucial calculus skill. Suppose you can choose the input of $f(x)$ and you want to pick the *best* value of x. The best value usually means the *maximum* value (if the function measures something desirable like profits) or the *minimum* value (if the function describes something undesirable like costs). We'll discuss the *optimization algorithm* in more detail in the next section, but first let us look at an example.

Crime TV

A calculus teacher turned screenwriter is working on the pilot episode for a new TV series. Here is the story he has written so far.

The local drug boss has recently been running into problems as police are intercepting his dealers on the street. The more drugs he sells, the more money he makes; but if he sells too much, police arrests will increase and he'll lose money. Fed up with this situation, he decides to find the *optimal* amount of drugs to release on the streets: as much as possible, but not enough to trigger the police raids. One day he tells his brothers and sisters in crime to leave the room and picks up a pencil and a piece of paper to do some calculus.

If x is the amount of drugs he releases on the street every day, then the amount of money he makes is given by the function

$$f(x) = 3000xe^{-0.25x},$$

where the linear part $3000x$ represents his profits with no police involvement and the $e^{-0.25x}$ represents the effects of the police stepping up their actions as more drugs are released.

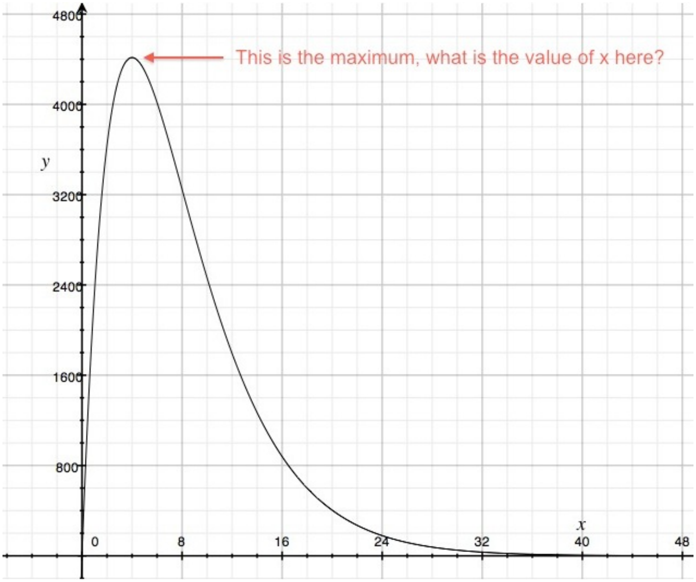


Figure 5.7: The graph of profit as a function of quantity sold.

Looking at the function, the drug boss asks, "What is the value of x which will give me the most profit from my criminal dealings?" Stated mathematically, he is asking,

$$\underset{x}{\operatorname{argmax}}\ 3000xe^{-0.25x} \ = ?$$

which means "find the value of the argument x that gives the *maximum* value of $f(x)$."

Remembering a conversation with a crooked financial analyst he met in prison, the drug boss recalls the steps required to find the maximum of a function. First he must take the function's derivative. Because the function is a product of two functions, he applies the product rule $[g(x)h(x)]' = g'(x)h(x) + g(x)h'(x)$. Taking the derivative of $f(x)$, he obtains

$$f'(x) = 3000e^{-0.25x} + 3000x(-0.25)e^{-0.25x}.$$

Whenever $f'(x) = 0$, the function $f(x)$ has zero slope. A maximum is exactly the kind of place where you'll find zero slope—think of a mountain peak with steep slopes on all sides; the mountain is momentarily horizontal at its peak.

So when is the derivative zero? We set up the equation,

$$f'(x) = 3000e^{-0.25x} + 3000x(-0.25)e^{-0.25x} = 0.$$

We factor out the 3000 and the exponential function to obtain

$$3000e^{-0.25x}(1 - 0.25x) = 0.$$

Since $3000 \neq 0$ and $e^{-0.25x} \neq 0$, the term in the bracket must be equal to zero:

$$(1 - 0.25x) = 0,$$

or $x = 4$. The slope of $f(x)$ is equal to zero when $x = 4$. This x value corresponds to the peak of the curve.

Then and there, the crime boss calls his posse back into the room and proudly announces that from then on, his organization will release exactly four kilograms of drugs per day.

"Boss, how much money will we make per day if we sell four kilograms?" asks one of the gangsters wearing sports pants.

"We'll make the *maximum* possible!" replies the boss.

"Yes I know Boss, but how much money is the maximum?"

The dude in sports pants is asking a good question. It is one thing to know *where* the maximum occurs, and it's another to know the value of the function at this point. The dude is asking the following mathematical question:

$$\max_x \ 3000xe^{-0.25x} \ = ?$$

Since we already know the maximum occurs at $x^* = 4$, we can plug this value into the function $f(x)$ to find

$$\max_x f(x) = f(4) = 3000(4)e^{-0.25(4)} = \frac{12000}{e} \approx 4414.55.$$

After this conversation is complete, everyone, including the boss, begins to question their choice of occupation in life. When you crunch the numbers, is crime really worth it?

A word of caution

It may seem funny to imagine calculus in the hands of the "bad guys," but in reality this is often the case. The System is obsessed with this whole optimization thing. Optimize to make more profits, optimize to minimize costs, optimize stealing of natural resources from Third World countries, optimize anything that moves, basically. Therefore, the System wants *you*—the young and powerful generation of the future—to learn this important skill and become faithful employees of corporations. The corporates want you to learn calculus so you can help them optimize things, ensuring the smooth continuation of the whole enterprise.

Mathematical knowledge does not come with an ethics manual to help you decide what should and should not be optimized; this responsibility falls on you. If, like me, you don't want to become a corporate sellout, you can always choose to use calculus for science. It doesn't matter whether it will be physics, medicine, or running your own company, it is all good. Just stay away from the System. Please do this for yourself and our future, will you?

Having said these words of warning, let's now proceed so I can show you the powerful optimization algorithm.

5.10 Optimization algorithm

This section shows and explains the details of the algorithm for finding the maximum of a function. This is called *optimization*, as in finding the optimal solution to a problem.

Say you have the function $f(x)$, which represents a real-world phenomenon. For example, $f(x)$ could represent how much *fun* you have as a function of alcohol consumed during one evening. We all know that with too much x, the fun stops and you find yourself, as the Irish say, "talking to God on the big white phone." Too little x and you might not have enough Dutch courage to chat up that girl/guy from the table across the room. To have as much fun as possible, you want to find the alcohol consumption x^* where f takes on its maximum value.

This is one of the prominent applications of calculus (I'm talking about optimization, not alcohol consumption). This is why you've been learning about all those limits, derivative formulas, and differentiation rules in previous sections.

Definitions

- x: the *variable* we can control
- $[x_i, x_f]$: the interval of values from which x can be chosen. The values of x must obey $x_i \leq x \leq x_f$. These are the *constraints* of the optimization problem. For the drinking optimization problem, $x \geq 0$ since you can't drink negative alcohol, and probably $x < 2$ (in litres of hard booze) because roughly at this point a person will die from alcohol poisoning. So we are searching for the optimal amount of alcohol x in the interval $[0, 2]$.

- $f(x)$: the *function* we want to optimize. This function must be *differentiable*, meaning we can take its derivative.
- $f'(x)$: the *derivative* of $f(x)$. The derivative contains information about the slope of $f(x)$.
- *maximum*: a place where the function reaches a peak. When there are multiple peaks, we call the highest peak the *global maximum*, while all other peaks are *local maxima*.
- *minimum*: a place where the function reaches a low point at the bottom of a valley. The *global minimum* is the lowest point overall, whereas a *local minimum* is the minimum in some neighbourhood.
- *extremum*: a general term to describe both maximum and minimum points.
- *saddle point*: a place where $f'(x) = 0$ at a point that is neither a max nor a min. For example, the function $f(x) = x^5$ has a saddle point at $x = 0$.

Suppose some function $f(x)$ has a global maximum at x^*, and the value of that maximum is $f(x^*) = M$. The following mathematical notations apply:

- $\text{argmax}_x\ f(x) = x^*$: the location (the *argument* of the function) where the maximum occurs
- $\max_x\ f(x) = M$: the maximum value

Algorithm for finding extrema

Input: a function $f(x)$ and a constraint region $C = [x_i, x_f]$
Output: the location and value of all maxima and minima of $f(x)$

Follow this algorithm step-by-step to find the extrema of a function:

1. First, *look* at $f(x)$. If you can plot it, plot it. If not, try to imagine what the function looks like.

2. Find the derivative $f'(x)$.

3. Solve the equation $f'(x) = 0$. Usually, there will be multiple solutions. Make a list of them. We'll call this the list of *candidates*.

4. For each candidate x^* on the list, check to see whether it is a max, a min, or a saddle point:

    - If $f'(x^* - 0.1)$ is positive and $f'(x^* + 0.1)$ is negative, then the point x^* is a max. The function goes up, flattens at x^*, then goes down after x^*. Therefore, x^* must be a peak.
    - If $f'(x^* - 0.1)$ is negative and $f'(x^* + 0.1)$ is positive, the point x^* is a min. The function goes down, flattens, then goes up, so the point must be a minimum.
    - If $f'(x^* - 0.1)$ and $f'(x^* + 0.1)$ have the same sign, the point x^* is a saddle point. Remove it from the list of candidates.

5. Now go through the list one more time and reject all candidates x^* that do not satisfy the constraints C. In other words, if $x \in [x_i, x_f]$, the candidate stays; but if $x \notin [x_i, x_f]$, we remove it since this solution is not *feasible*. Returning to the alcohol consumption example, if you have a candidate solution that says you should drink 5[L] of booze, you must reject it because otherwise you would die.

6. Add x_i and x_f to the list of candidates. These are the boundaries of the constraint region and should also be considered. If no constraint was specified, use the *default* constraint region $-\infty < x < \infty$ and add $-\infty$ and ∞ to the list of candidates.

7. For each candidate x^*, calculate the function value $f(x^*)$.

The resulting list is a collection of *local* extrema: maxima, minima, and endpoints. The *global maximum* is the largest value from the list of local maxima. The *global minimum* is the smallest of the local minima.

Note that in dealing with points at infinity, such as $x^* = \infty$, we don't actually calculate a value; rather, we calculate the limit $\lim_{x\to\infty} f(x)$. Usually, the function either blows up $f(\infty) = \infty$ (like x, x^2, e^x, $\ldots$), drops down indefinitely $f(\infty) = -\infty$ (like $-x$, $-x^2$, $-e^x$, $\ldots$), or reaches some value (like $\lim_{x\to\infty} \frac{1}{x} = 0$, $\lim_{x\to\infty} e^{-x} = 0$). If a function goes to positive ∞ it doesn't have a global maximum and instead continues growing indefinitely. Similarly, functions that go toward negative ∞ don't have a global minimum.

Example 1 Find all the maxima and minima of the function

$$f(x) = x^4 - 8x^2 + 356.$$

Since no interval is specified, we'll use the default interval $x \in \mathbb{R}$. Let's go through the steps of the algorithm.

1. We don't know what the x^4 function looks like, but it is probably similar to the x^2—it goes up to infinity on the far left and the far right.

2. Using the formula for derivative of polynomials we find

$$f'(x) = 4x^3 - 16x.$$

3. Now we must solve

$$4x^3 - 16x = 0,$$

which is the same as

$$4x(x^2 - 4) = 0,$$

which is the same as

$$4x(x-2)(x+2) = 0.$$

The list of candidate points is $\{x = -2, x = 0, x = 2\}$.

4. For each of these points, we'll check to see if it is a max, a min, or a saddle point.

(a) For $x = -2$, we check $f'(-2.1) = 4(-2.1)(-2.1-2)(-2.1+2) < 0$ and $f'(-1.9) = 4(-1.9)(-1.9-2)(-1.9+2) > 0$ to conclude $x = -2$ must be a minimum.

(b) For $x = 0$ we try $f'(-0.1) = 4(-0.1)(-0.1-2)(-0.1+2) > 0$ and $f'(0.1) = 4(0.1)(0.1-2)(0.1+2) < 0$, which reveals we have a maximum at $x = 0$.

(c) For $x = 2$, we check $f'(1.9) = 4(1.9)(1.9-2)(1.9+2) < 0$ and $f'(2.1) = 4(2.1)(2.1-2)(2.1+2) > 0$, so $x = 2$ must be a minimum.

5. We don't have any constraints, so all of the above candidates make the cut.

6. We add the two constraint boundaries $-\infty$ and ∞ to the list of candidates. At this point, our final shortlist of candidates contains $\{x = -\infty, x = -2, x = 0, x = 2, x = \infty\}$.

7. We now evaluate the function $f(x)$ for each of the values to obtain location-value pairs $(x, f(x))$, like so: $\{(-\infty, \infty)$, $(-2, 340)$, $(0, 356)$, $(2, 340)$, $(\infty, \infty)\}$. Note that $f(\infty) = \lim_{x\to\infty} f(x) = \infty^4 - 8\infty^2 + 356 = \infty$ and the same is true for $f(-\infty) = \infty$.

We are done. The function has no global maximum since it increases to infinity. It has a local maximum at $x = 0$ with value 356. It also has two global minima at $x = -2$ and $x = 2$, both of which have value 340. Thank you, come again.

Alternate algorithm

Instead of checking nearby points to the left and right of each critical point, we can modify the algorithm with an alternate Step 4 known as the *second derivative test*. Recall the second derivative tells us the function's *curvature*. If the second derivative is positive at a critical point x^*, then the point x^* must be a minimum. If, on the other hand, the second derivative at a critical point is negative, the function must be maximum at x^*. If the second derivative is zero, the test is inconclusive.

Alternate Step 4

- Check each candidate x^* to determine if is a max, a min, or a saddle point.
  - ▷ If $f''(x^*) < 0$ then x^* is a max.
  - ▷ If $f''(x^*) > 0$ then x^* is a min.
  - ▷ If $f''(x^*) = 0$ then the second derivative test fails. We must revert back to checking nearby values $f'(x^* - \delta)$ and $f'(x^* + \delta)$ to determine if x^* is a max, a min, or a saddle point.

Limitations

The optimization algorithm above applies to *differentiable* functions of a single variable. Not all functions are differentiable. Functions with sharp corners, such as the absolute value function $|x|$, are not differentiable everywhere, and therefore won't work with the algorithm above. Functions with jumps in them, like the Heaviside step function, are not continuous and therefore not differentiable—the algorithm cannot be used on them either.

We can generalize the optimization procedure, which help us optimize functions of multiple variables $f(x, y)$. You'll learn how to do this in the course *multivariable calculus*. The optimization techniques will be similar to the steps above, but with more variables and more intricate constraint regions.

At last, I want to comment on the fact that you can only maximize *one* function. Say the drug boss from the TV series wanted to maximize his funds $f(x)$ *and* his gangster street cred $g(x)$. This is not a well-posed problem; either you maximize $f(x)$ or you maximize $g(x)$, but you can't do both. There is no reason why a single x would give the highest value for both $f(x)$ and $g(x)$. If both functions are important to you, you can make a new function that combines the original two $F(x) = f(x) + g(x)$ and maximize $F(x)$. If gangster street cred is three times more important to you than funds, you could optimize $F(x) = f(x) + 3g(x)$, but it is mathematically and logically impossible to maximize two things at the same time.

Exercises

The function $f(x) = x^3 - 2x^2 + x$ has a local maximum on the interval $x \in [0, 1]$. Find where this maximum occurs, and find the value of f at that point.
Ans: $\left(\frac{1}{3}, \frac{4}{27}\right)$.

5.11 Implicit differentiation

Thus far, we've discussed how to compute derivatives of functions $f(x)$. When we identify the function's output with the variable y, we can write $y(x) = f(x)$, which shows the variable y depends on x through the function $f(x)$. The slope of this function is calculated as the rise in the y-direction divided by the run in the x-direction $y'(x) \equiv \frac{dy}{dx} \equiv f'(x)$.

We can also use the derivative operation to compute the slope in *mathematical relations* that are not expressed in the form $y(x) = f(x)$. For example, consider the equation that describes a circle of radius R:

$$x^2 + y^2 = R^2.$$

The equation of a circle describes a *relation* between the variables x and y, without specifying one variable as a function of the other. Nevertheless, we can still treat y as a function of x. We say the function $y(x)$ is *implicit*.

If we want to make the functional relationship between y and x *explicit*, we can rewrite the equation $x^2 + (y(x))^2 = R^2$ in the form

$$y(x) = \pm\sqrt{R^2 - x^2},$$

which shows *explicitly* how y depends on the variable x.

Problem Consider the point $P = (x_P, y_P)$, which lies on the circle $x^2 + y^2 = R$. Find the *slope* of the tangent line to the circle at that point.

This problem is asking us to find $y'(x_P)$. Using the explicit function $y(x)$, we would first compute the derivative function $y'(x) = \pm\frac{1}{2}\frac{1}{\sqrt{R^2-x^2}}(-2x)$ and then substitute the value x_P into $y'(x)$. The slope of the tangent line to the circle at the point $P = (x_P, y_P)$ is

$$y'(x_P) = \frac{-x_P}{\sqrt{R^2 - x_P^2}} = -\frac{x_P}{y_P}\,.$$

But do we really need to go through the explicit equation? Let me show you a faster way to answer the problem, without using the explicit function $y(x)$. Start by taking the derivative of the equation that describes the circle:

$$\begin{aligned} \frac{d}{dx}\big[\ x^2\ \ +\ \ \ y^2\ \ &= R^2\big], \\ 2x + 2y\frac{dy}{dx} &= 0, \\ \frac{dy}{dx} &= -\frac{x}{y}\,. \end{aligned}$$

The slope at $P = (x_P, y_P)$ is therefore $\dfrac{dy}{dx} = -\dfrac{x_P}{y_P}\,.$

Note how we used the chain rule for the implicit function $y(x)$.

Definitions

- $g(x, y) = 0$: a *relation* between the variables x and y
- $\frac{d}{dx}\big[g(x, y)\big] = 0$: the derivative of the relation $g(x, y) = 0$ with respect to the variable x
- $dg = \dfrac{dg}{dx}dx + \dfrac{dg}{dy}dy$: the *total derivative* of the function $g(x, y)$

Explanations

The equation of a circle can be written as a function $g(x, y)$ as

$$g(x, y) \equiv x^2 + y^2 - R^2 = 0.$$

The *implicit* derivative of this equation with respect to x is

$$\begin{aligned}\frac{d}{dx}[g(x,y)] = \frac{d}{dx}\left[x^2+y^2-R^2\right] &= \frac{d}{dx}[0]\\ \frac{dg}{dx} + \frac{dg}{dy}\frac{dy}{dx} &= 0\\ 2x \ + 2y\frac{dy}{dx} &= 0\end{aligned}$$

What is *implicit* in this derivative calculation is the assumption that y is a function of x. The expression $\frac{dy}{dx}$ refers to the derivative of the implicit function $y(x)$. After isolating $\frac{dy}{dx} \equiv y'(x) = \frac{-x}{y}$, we are able to find the slope for any point $P = (x_P, y_P)$ on the circle. You can check that the slope predicted for $P = (0, R)$ is 0. Also check that the slope is infinite at $P = (R, 0)$, since the tangent to the circle is vertical.

Total derivative

Consider some relation $g(x,y) = 0$ and assume that both x and y are implicit functions of some third variable t. If we compute the derivative of the expression $g(x,y)$ with respect to t we obtain

$$\frac{dg}{dt} = \frac{dg}{dx}\frac{dx}{dt} + \frac{dg}{dy}\frac{dy}{dt}\,.$$

We call this the *total* derivative of g because we computed the dependence between g and t through both functions $x(t)$ and $y(t)$. Note each term in the total derivative is obtained through the chain rule:

$$\frac{d}{dt}g(x(t),y) = \frac{dg}{dx}\frac{dx}{dt} \quad \text{and} \quad \frac{d}{dt}g(x,y(t)) = \frac{dg}{dy}\frac{dy}{dt}.$$

Let's look at an example that involves the total derivative.

Example In the corporate world, a man's ego E is related to his salary S by the following equation:

$$E^2 = S^3.$$

Both E and S are functions of time. What is the rate of change of the ego of Corporate Joe, the insurance analyst, when he makes 60k and his salary increases at a rate of 5k per year?

This is called a *related rates* problem. We are told that $\frac{dS}{dt} = 5000$ and we're asked to find $\frac{dE}{dt}$ when $S = 60000$. First, take the implicit derivative of the salary-to-ego relation:

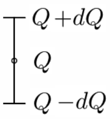
$$\begin{aligned} \frac{d}{dt}\left[\, E^2 \,\right] &= \frac{d}{dt}\left[\, S^3 \,\right], \\ 2E\frac{dE}{dt} &= 3S^2\frac{dS}{dt}. \end{aligned}$$

We are interested in the point where $S = 60000$. To find Joe's ego at this point, apply the original relation $E^2 = S^3$ and solve for E to find $E = \sqrt{60000^3} = 14696938.46$ ego points. Substituting all these values into the derivative of the relation, we find

$$2(14696938.46)\frac{dE}{dt} = 3(60000)^2(5000).$$

Joe's ego is growing at $\frac{dE}{dt} = \frac{3(60000)^2(5000)}{2(14696938.46)} = 1837117.31$ ego points per year. Yay, ego points! I wonder what you can redeem these for.

Error bars



In science, when we report the results of an experimental measurement of some quantity Q, we write $Q \pm dQ$, where dQ is an estimate of the error of the measurement. The measurement error dQ is represented graphically as an "error bar" as shown on the right. The *precision* of a measurement is defined as the *ratio* of the error of the measurement divided by the size of the quantity being measured $\frac{dQ}{Q}$, or as a percentage.

Suppose the quantity C depends on the variables x and y. We can express the dependence between the error in the measurement of C and the error in the measurement of x and y using the formula

$$dC = \frac{dC}{dx}dx + \frac{dC}{dy}dy.$$

Note the similarity to the formula for the total derivative.

Example You want to calculate the kinetic energy of a particle. Recall the formula for kinetic energy, $K = \frac{1}{2}mv^2$. Suppose you measured the particle's mass m with precision 3%, and the particle's velocity with precision 2%. What will be the precision of your kinetic energy calculation?

We want to find $\frac{dK}{K}$ and are told $\frac{dm}{m} = 0.03$ and $\frac{dv}{v} = 0.02$. The first step is to calculate the *implicit derivative* of the expression for the kinetic energy:

$$dK = d\left(\frac{1}{2}mv^2\right) = \frac{dK}{dm}dm + \frac{dK}{dv}dv = \frac{1}{2}v^2(dm) + mv(dv),$$

in which we apply the product rule of derivatives. To obtain the relative error, divide both sides by K to obtain

$$\frac{dK}{K} = \frac{\frac{1}{2}v^2\,dm \;+\; m\,v\,dv}{\frac{1}{2}m\,v^2} = \frac{dm}{m} + 2\frac{dv}{v}.$$

The result tells us the precision of the kinetic energy measurement is $\frac{dK}{K} = 0.03 + 2(0.02) = 0.07$ or 7%. Note the error in the velocity measurement dv contributes twice as much as the error in the mass measurement dm to the error dK. This is because the velocity appears with exponent 2 in the formula $K = \frac{1}{2}mv^2$.

Discussion

We have reached the half-point of the calculus chapter. We learned what derivatives are, and we described applications of derivatives for optimization problems, finding tangent lines, and computing related rates.

Before you continue reading about integrals, I highly recommend you attempt to solve some of the exercises on page 409 in the back of the chapter. Understanding the theory is important, but it is by solving exercises that you will become a calculus expert.

5.12 Integrals

We now begin our discussion of integrals, the second topic in calculus. An integral is a fancy way of computing the area under the graph of a function. Integral calculus is usually taught as a separate course after differential calculus, but this separation can be counter-productive. The easiest way to understand integration is to think of it as the inverse of the derivative operation. Integrals are antiderivatives. Once you realize this fundamental fact, you'll be able to apply all your differential calculus knowledge to the domain of integral calculus. In differential calculus, we learned how to take a function $f(x)$ and find its derivative $f'(x)$. In integral calculus, we will be given a function $f(x)$ and we'll be asked to find its *antiderivative* function $F(x)$. The antiderivative of $f(x)$ is a function $F(x)$ whose derivative equals $f(x)$.

In this section, we'll learn about two tasks: how to compute antiderivatives, and how to compute the area under the graph of $f(x)$. Confusingly, both of these tasks are called *integration*. To avoid any possibility of confusion, let's define things clearly:

- The *indefinite integral* of $f(x)$ is denoted $\int f(x)dx = F(x) + C$. To compute the indefinite integral of $f(x)$, you must find a function $F : \mathbb{R} \to \mathbb{R}$, such that $F'(x) = f(x)$. The indefinite integral is the antiderivative function.
- The *definite integral* of $f(x)$ between $x = a$ and $x = b$ is denoted $\int_a^b f(x)dx = A(a, b)$. Definite integrals correspond to the computation of the area under the function $f(x)$ between $x = a$ and $x = b$. The definite integral is a number $A(a, b) \in \mathbb{R}$.

The two integration tasks are related. The area under the curve $A(a, b)$ can be computed as the *change* in the antiderivative function, using to the formula $A(a, b) = \big[F(x) + C\big]_a^b = F(b) - F(a)$.

Now let's look at the details.

Definitions

You should already be familiar with these concepts:

- $\mathbb{R}$: the set of real numbers
- $f(x)$: a function of the form $f : \mathbb{R} \to \mathbb{R}$, which means f takes real numbers as inputs and produces real numbers as outputs
- $\lim_{\delta \to 0}$: a limit expression in which the number δ tends to zero
- $f'(x)$: the derivative of $f(x)$ is the rate of change of f at x:

  $$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) \ - \ f(x)}{\delta}.$$

  The derivative is a function of the form $f' : \mathbb{R} \to \mathbb{R}$.

These are the new concepts, which we will learn about in integral calculus:

- $A(a,b)$: the value of the *area* under the curve $f(x)$ from $x = a$ until $x = b$. The area $A(a,b)$ is computed as the following integral

  $$A(a,b) = \int_a^b f(x)\, dx.$$

  The $\int$ sign stands for *sum*. Indeed, the integral is the "sum" of $f(x)$ for all values of x between a and b.

- $A_0(x)$: the *integral function* of $f(x)$. The integral function corresponds to the computation of the area under $f(x)$ as a function of the upper limit of integration:

  $$A_0(x) \equiv A(0,x) = \int_0^x f(u)\, du.$$

  The choice of $x = 0$ as the lower limit of integration is arbitrary.

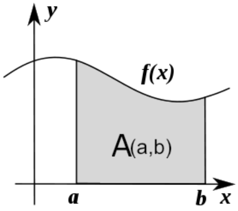
- $F(x)+C$: The *antiderivative* function of the function $f(x)$. An antiderivative function is defined as a function whose derivative equals to $f(x)$. The antiderivative function always includes an additive constant C. If the function $F(x)$ is an antiderivative (obeys $F'(x)=f(x)$) then the function $F(x)+C$ is also an antiderivative since
$$\frac{d}{dx}[F(x)+C]=f(x),$$
for any constant C.

- The fundamental theorem of calculus (FTC) states that the integral function $A_0(x)$ is equal to the antiderivative function $F(x)$ up to an additive constant C:
$$A(0,x)\equiv A_0(x)\overset{\text{FTC}}{=}F(x)+C.$$
The fundamental theorem leads us to the following formula for computing the area $A(a,b)$:
$$A(a,b)=A(0,b)-A(0,a)=A_0(b)-A_0(a)=F(b)-F(a).$$
The area under the curve, $A(a,b)$, is equal to the change in $F(x)$ between $x=a$ and $x=b$.

The area under the curve

An integral describes the computation of the area under the curve $f(x)$ between $x=a$ and $x=b$:

$$A(a,b)\equiv\int_a^b f(x)\,dx.$$



We refer to the numbers a and b as the *limits of integration*. The location where the integral starts, $x=a$, is called the *lower limit* of integration. The location where the integral stops, $x=b$, is called the *upper limit* of integration.

The integral as a function

The *integral function* of $f(x)$ describes the "running total" of the area under the curve $f(x)$ as a function of the upper limit of integration:

$$A_0(x) \equiv A(0,x) \equiv \int_0^x f(u)\,du.$$

The variable x represents the upper limit of integration. The variable u inside the integral is called the *integration variable* and its value varies between $u=0$ and $u=x$. The name of the integration variable u is not important; we can write $\int_0^x f(y)dy$ or $\int_0^x f(z)dz$ or even $\int_0^x f(\xi)d\xi$ and all of these represent the same function $A_0(x)$.

The choice of the lower limit of integration is also not important. For the sake of concreteness, we define the integral function to start at $x=0$. A different choice for the lower limit of integration would lead to a different integral function. For example, the integral function that describes the area under $f(x)$ starting from $x=a$ is defined as $A_a(x) \equiv A(a,x) \equiv \int_a^x f(u)du$. The function $A_0(x)$ can be obtained from the function $A_a(x)$ by adding the missing area $A(0,a)$:

$$\begin{aligned}\int_0^x f(u)\,du \; &= \int_0^a f(v)\,dv + \int_a^x f(w)\,dw \\ A(0,x) \; &= A(0,a) \qquad + \;\; A(a,x) \\ A_0(x) \; &= A_0(a) \qquad + \;\; A_a(x).\end{aligned}$$

The area $A(a,b) \equiv A_a(b)$ can be computed as the *change* in the value of $A_0(x)$ between $x=a$ and $x=b$:

$$A(a,b) \equiv \int_a^b f(x)\,dx = A_0(b) - A_0(a).$$

Note the formula $A(a,b) = A_c(b) - A_c(a)$ applies for all $c \in \mathbb{R}$.

The antiderivative function

The antiderivative function $F(x)$ of $f(x)$ is a function whose derivative equals $f(x)$:

$$\frac{d}{dx}\left[\ F(x)\ \right] = f(x).$$

The antiderivative function is not unique: any function $F(x) + C$ also obeys $\frac{d}{dx}\big[F(x) + C\big] = f(x)$, since the derivative of a constant is zero.

The fundamental theorem of calculus

The fundamental theorem of calculus states that the integral function $A_c(x)$ is an antiderivative of $f(x)$:

$$\frac{d}{dx}\big[A_c(x)\big] = \frac{d}{dx}\int_c^x f(u)\,du = f(x).$$

Thus far, we spoke of integral functions $A_c(x)$ and antiderivative functions $F(x)+C$ as different mathematical objects, but the fundamental theorem of calculus implies the equation

$$A_c(x) = F(x) + C.$$

Every integral function $A_c(x)$ is also an antiderivative function, and every antiderivative function $F(x) + C$ corresponds to the integral function $A_c(x)$, for some $c \in \mathbb{R}$. From this point on, we will use the notation $F(x)$ to refer to the integral function.

We'll discuss the fundamental theorem of calculus in more detail when we reach Section 5.14. For now, let's focus on the task of computing integrals by performing reverse-differentiation.

Indefinite integrals

The function $F(x)$ is the result of applying the integral *operator* $\int \cdot\, dx$ to the function $f(x)$. The integral operator $\int \cdot\, dx$ takes functions as inputs and produces functions as outputs:

$$\int \cdot \, dx : \{\mathbb{R} \to \mathbb{R}\} \quad \to \quad \{\mathbb{R} \to \mathbb{R}\}.$$

The integral operator takes a function $f(x)$ as an input and produces its antiderivative function $F(x)$ as output. Like the derivative function $f'(x)$, the integral function $\int f(x)\,dx$ describes a *property* of the original function $f(x)$. We use the derivative operator to find the "slope of the graph" property of a function. We use the integral operator to find the "area under the graph" property of a function.

In an *indefinite* integral problem, we are given a function $f(x)$ and asked to find its integral function $F(x)$:

$$F(x) = \int f(x)\,dx.$$

The integral is *indefinite* because we're performing an integral calculation but haven't defined the limits of integration $x = a$ and $x = b$.

As a consequence of the fundamental theorem of calculus, we know the derivative of the function $F(x)$ is equal to $f(x)$:

$$F'(x) = f(x).$$

Thus, to find an integral function of the function $f(x)$, we must find a function $F(x)$ such that $F'(x) = f(x)$.

Example Suppose you want to find the indefinite integral $\int x^2\,dx$. Using the fundamental theorem, we can rephrase this problem as the search for some function $F(x)$ such that

$$F'(x) = x^2.$$

Since you remember your derivative formulas, you can guess right away that $F(x)$ must contain an x^3 term. From the power rule for derivatives, you know that taking the derivative of a cubic term results in a quadratic term. Therefore, the function you are looking for has the form $F(x) = cx^3$, for some constant c. Pick the constant c that makes this equation true:

$$F'(x) = 3cx^2 = x^2.$$

Solving $3c = 1$, we find $c = \frac{1}{3}$ and so the answer to this indefinite integral problem is

$$\int x^2 \, dx = \frac{1}{3}x^3 + C.$$

You can verify that $\frac{d}{dx}\left[\frac{1}{3}x^3 + C\right] = x^2$. Did you see what just happened? We were able to take an integral using only derivative formulas and "reverse engineering."

Example 2 Since we know

$$F(x) = x^4 \qquad \xrightarrow{\frac{d}{dx}} \qquad F'(x) = 4x^3 \equiv f(x),$$

then it must be that

$$f(x) = 4x^3 \qquad \xrightarrow{\int dx} \qquad F(x) = \int 4x^3 \, dx = x^4 + C.$$

Example 3 Let's look at some more integrals:

- The indefinite integral of $f(x) = \cos\theta$ is

$$F(x) = \int \cos\theta \; d\theta = \sin\theta + C,$$

  since $\frac{d}{d\theta}\sin\theta = \cos\theta$.

- Similarly, the integral of $f(x) = \sin\theta$ is

$$F(x) = \int \sin\theta \; d\theta = -\cos\theta + C,$$

  since $\frac{d}{d\theta}[-\cos\theta] = \sin\theta$.

- The integral of $f(x) = x^n$ for any number $n \neq -1$ is

$$F(x) = \int x^n \; dx = \frac{1}{n+1} x^{n+1} + C,$$

since $\frac{d}{d\theta} x^n = n x^{n-1}$.

- The integral of $f(x) = x^{-1} = \frac{1}{x}$ is

$$F(x) = \int \frac{1}{x} \; dx = \ln x + C,$$

since $\frac{d}{dx} \ln x = \frac{1}{x}$.

I could go on but I think you get the point: all the derivative formulas you learned (see page 301) can be used in the opposite direction as integral formulas.

Remember to always add a constant term $+C$ to your answer. The answer to the indefinite integral question $\int f(x)\, dx$ is not a single function $F(x)$, but a whole family of functions $F(x) + C$ that differ by an additive constant C.

Definite integrals

A *definite* integral specifies the function to integrate as well as the limits of integration $x = a$ and $x = b$. The area under $f(x)$ between $x = a$ and $x = b$ is

$$A(a, b) \equiv \int_a^b f(x) \; dx.$$

To find the value of the definite integral, we will proceed in two steps. The first step is to calculate the antiderivative function $\int f(x) \; dx = F(x) + C$. In other words, you must solve the indefinite integral before you solve the definite integral.

The second step is to compute the area $A(a, b)$ as the change in the antiderivative function between $x = a$ and $x = b$:

$$A(a, b) = \big[F(x) + C\big]\Big|_{x=a}^{x=b} = [F(b) + C] - [F(a) + C] = F(b) - F(a).$$

Note the new "vertical bar" notation: $g(x)\Big|_{\alpha}^{\beta} = g(\beta) - g(\alpha)$. This is a useful shorthand for denoting the change in the function $g(x)$ between two points. Figure 5.8 illustrates the meaning of this procedure.

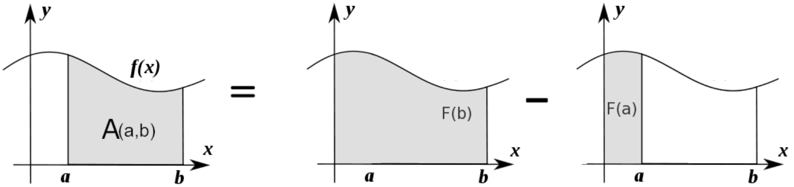


Figure 5.8: The function $F(x)$ measures the area under the curve $f(x)$. The area under $f(x)$ between $x = a$ and $x = b$ is $A(a, b) = F(b) - F(a)$.

You can also try rearranging the plots in Figure 5.8 to visualize the equation $A(0, b) = A(0, a) + A(a, b)$. The "running total" of the area under $f(x)$ until $x = b$ is equal to the "running total" of the area under $f(x)$ until $x = a$, plus the area $A(a, b)$. The formula $A(a, b) = F(b) - F(a)$ follows from combining the equation $A(0, b) = A(0, a) + A(a, b)$ with the result of the fundamental theorem of calculus: $A(0, x) \equiv A_0(x) \overset{\text{FTC}}{=} F(x)$.

Example 4 The antiderivative of $f(x) = x^2$ is $F(x) = \frac{1}{3}x^3 + C$. Use this fact to find the value of the definite integral $\int_a^b x^2\, dx$.

The definite integral is computed by evaluating the value of the antiderivative function at the upper limit and subtracting the value of the antiderivative function at the lower limit:

$$\int_a^b x^2\, dx = \left[\tfrac{1}{3}x^3 + C\right]_{x=a}^{x=b} = \left[\tfrac{1}{3}b^3 + C\right] - \left[\tfrac{1}{3}a^3 + C\right] = \frac{1}{3}(b^3 - a^3).$$

Example 5 What is the area under the curve $f(x) = \sin(x)$, between $x = 0$ and $x = \pi$? First we take the antiderivative

$$F(x) = \int \sin(x)\; dx = -\cos(x) + C.$$

Now we calculate the difference between $F(x)$ at the upper limit minus $F(x)$ at the lower limit:

$$\begin{aligned}
A(0,\pi) &= \int_0^{\pi} \sin(x)\; dx \\
&= \underbrace{[-\cos(x) + C]}_{F(x)}\Big|_0^{\pi} \\
&= [-\cos\pi + C] - [-\cos(0) + C] \\
&= \cos(0) - \cos\pi \quad = \quad 1 - (-1) = 2.
\end{aligned}$$

The final answer does not depend on the constant C because we evaluate the *change* in $F(x) + C$ and so C cancels out.

In case you are wondering what the "area under the curve" calculation is used for in practice, you should recall how we derived the kinematics equations in Chapter 2. The velocity $v(t)$ measures change in position $x(t)$ over time. The total change in position between $t = 0$ and $t = \tau$ is obtained by calculating the integral of $v(t)$ as follows:

$$x(\tau) - x(0) = \int_0^{\tau} v(t)\; dt.$$

Note how the dimensions work in this equation. Time is measured in seconds [s], and $v(t)$ is measured in [m/s], so the area under $v(t)$ has dimensions of [m/s]×[s] = [m].

Properties of integrals

Signed area

The value of a definite integral can be either positive or negative. If the limits of integration a and b satisfy $a < b$ (b is to the right of a on the number line), and if $f(x) > 0$ (meaning $f(x)$ is a positive function), then the area under the curve will be positive:

$$A(a,b) = \int_a^b f(x)\,dx \quad > \quad 0.$$

For a function $g(x) < 0$, the integral from a to b corresponds to a negative area. In general, if $f(x)$ is above the x-axis in some places, these zones will contribute positively to the total area under the curve; places where $f(x)$ is below the x-axis will contribute negatively to the total area $A(a,b)$.

We can also obtain a negative area if we swap the limits of integration. Suppose we have $f(x) > 0$, and limits of integration a and b such that $a < b$. If we start integrating at $x = b$ and stop integrating at $x = a$, the area under the curve will be negative:

$$A(b,a) = \int_b^a f(x)\,dx \quad < \quad 0.$$

The function $f(x)$ is positive, but each integration step dx is *negative*, since we are moving from right to left.

Integrals are *signed* areas. Changing the direction of integration changes the sign of the integral:

$$A(b,a) = \int_b^a f(x)\;dx = -\int_a^b f(x)\;dx = -A(a,b).$$

Additivity

The integral from a to b plus the integral from b to c is equal to the integral from a to c:

$$A(a,b) + A(b,c) = \int_a^b f(x)\;dx + \int_b^c f(x)\;dx = \int_a^c f(x)\;dx = A(a,c).$$

Linearity

Integration is a linear operation, meaning

$$\int [\alpha f(x) + \beta g(x)]\;dx = \alpha \int f(x)\;dx + \beta \int g(x)\;dx$$

for arbitrary constants α, β. Recall the derivative is also a linear operation:

$$[\alpha f(x) + \beta g(x)]' = \alpha f'(x) + \beta g'(x).$$

Thus, we can say the operations of calculus as a whole are *linear* operations. This property is really cool, because it allows us to break down complicated problems into smaller chunks.

Explanations

In this section, we introduced several new concepts: the area under the curve $A(a, b)$, the integral function $A_c(x)$, and the antiderivative function $F(x) + C$. We also stated the fundamental theorem of calculus, which relates these three concepts:

$$A(c, x) \equiv A_c(x) \overset{\text{FTC}}{=} F(x) + C,$$

for some choice of the constants c and C. Let's define more precisely the equivalence between these concepts.

The set of antiderivative functions

The antiderivative function $F(x)+C$ always includes an arbitrary additive constant C. Thus it would be wrong to talk about *the* antiderivative as a single function; there is a whole set of functions that are antiderivatives of $f(x)$.

Let $\{\mathbb{R} \to \mathbb{R}\}$ denote the set of all functions that take real numbers as inputs and produce real numbers as outputs. The *set* of antiderivative functions for a function $f(x)$ is defined as

$$\left\{ F \in \{\mathbb{R} \to \mathbb{R}\} \;\middle|\; F'(x) = f(x) \right\}.$$

In words, the set of antiderivatives of the function $f(x)$ is the subset of all functions $\{\mathbb{R} \to \mathbb{R}\}$ such that $F'(x) = f(x)$.

Similarly, there is a whole set of integral functions, which differ by the choice of the starting point of integration:

$$\left\{ A_c \in \{\mathbb{R} \to \mathbb{R}\}, c \in \mathbb{R} \;\middle|\; A_c(x) = \int_c^x f(u)\,du \right\}.$$

The fundamental theorem of calculus states that the *set* of antiderivative functions is equal to the *set* of integral functions.

Observe that *any* function from the set of integral functions (antiderivatives) can be used for the area calculation:

$$A(a,b) \equiv \int_a^b f(x)\,dx = A_c(b) - A_c(a) = F(b) - F(a).$$

Integration is the inverse operation of differentiation

You are already familiar with the inverse relationship between *functions* from Chapter 1. You know that to solve for x in the equation $f(x) = c$, you must apply the inverse function f^{-1} to both sides of the equation. The function f^{-1} will undo the effects of f, leaving $x = f^{-1}(c)$.

There exists an analogous inverse relationship between the derivative operator $\frac{d}{dx}$ and the integral operator $\int \cdot\, dx$:

$$\text{int}(\text{diff}(F(x))) = \int_0^x \left(\frac{d}{du}F(u)\right) du = \int_0^x f(u)\,du \overset{\text{FTC}}{=} F(x) + C.$$

The integral operator $\int \cdot\, dx$ is the “undo button” for the derivative operator.

Applications

Suppose you want to find the function $f(t)$ that satisfies the differential equation

$$\frac{d}{dt}\Big[f(t)\Big] = 100.$$

To find $f(t)$ you must *undo* the $\frac{d}{dt}$ operation. After applying the integration operation to both sides of the equation, we obtain

$$\begin{aligned} \int \left(\frac{d}{dt}\, f(t) \right) dt &= \int (100)\ dt \\ f(t) &= 100t + C. \end{aligned}$$

The solution to the equation $f'(t) = 100$ is $f(t) = 100t + C$, where C is called the *integration constant*.

5.13 Riemann sums

Our discussion in the previous section focussed on the inverse relationship between the integral operator $\int f(x)dx$ and the derivative operator $\frac{d}{dx}$. We learned the antiderivative function $F(x)$ can be used to compute the area under a curve $f(x)$ using the formula $\int_a^b f(x)\, dx = F(b) - F(a)$. Thus, with your differentiation skills and some reverse engineering, you can now handle integrals too.

Is there a way to compute integrals without referring to the derivative operation? No course on calculus would be complete without telling the classic "rectangles story" for computing definite integrals, which goes by the name *Riemann sum*. The Riemann sum is a procedure for computing the area under a curve by breaking up the area into many, little rectangular strips with heights that vary according to $f(x)$. To obtain the total area under the curve, we sum all the areas of these little rectangles.

First, like a cast of characters, we'll introduce some definitions.

Definitions

- $f(x)$: a function $f\colon \mathbb{R} \to \mathbb{R}$
- a: where the sum starts
- b: where the sum stops

- $A(a, b)$: the exact value of the area under the curve $f(x)$ from $x = a$ until $x = b$
- $S_n(a, b)$: an approximation to the area A in terms of n rectangles
- s_k: the area of the k^{th} rectangle when counting from the left

In this section we will consider the calculation of the area under the curve $f(x) = x^3 - 5x^2 + x + 10$ between $x = -1$ and $x = 4$. Figure 5.9 shows the graph of $f(x)$ and an approximation of the area under the curve as the sum of the areas of 12 rectangles.

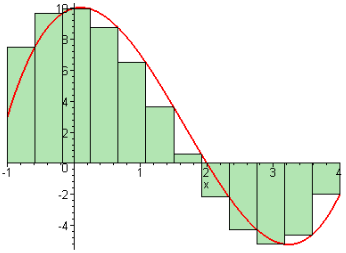


Figure 5.9: An approximation of the area under the function $f(x) = x^3 - 5x^2 + x + 10$ between $x = -1$ and $x = 4$ using $n = 12$ rectangles.

Formulas

The combined-area approximation is given by the *sum* of the areas of the little rectangles:

$$S_n(a, b) = \sum_{k=1}^{n} s_k.$$

Each of the little rectangles has an area s_k given by the rectangle's *height* multiplied by its *width*. The height of each rectangle will vary, but the width is constant. Why constant? Riemann figured that having each rectangle with a constant width Δx would make it easy to calculate the approximation. The total length of the interval from $x = a$ to $x = b$ is $(b - a)$. When we divide this length into n equally spaced segments, each segment will have width Δx given by

$$\Delta x = \frac{b-a}{n}.$$

Okay, we have the width formula; now let's find the height of the k^{th} rectangle in the sequence of rectangles. For the rectangles, we pick isolated "samples" of $f(x)$ for the following values:

$$x_k = a + k\Delta x, \text{ for } k \in \{1, 2, 3, \ldots, n\},$$

with all rectangles equally spaced $\Delta x = \frac{b-a}{n}$ apart.

The function's height varies as we move along the x-axis. The area of each rectangle is equal to its height $f(x_k)$ times its width:

$$s_k = f(a + k\Delta x)\Delta x.$$

Now, my dear students, I want you to stare at the above equation and do some simple calculations to check that you understand. There is no point in continuing if you are just taking my word for it. **Verify** that when $k = 1$, the formula gives the area of the first little rectangle. **Verify also** that when $k = n$, the formula $x_k = a + k\Delta x$ reaches the upper limit b.

Let's put our formula for s_k in the sum where it belongs. The Riemann sum approximation using n rectangles is given by

$$S_n(a, b) = \sum_{k=1}^{n} f(a + k\Delta x)\Delta x,$$

where $\Delta x = \frac{b-a}{n}$.

The integral is defined as the limit of the Riemann sum as n goes to infinity: $\int_a^b f(x)\,dx \equiv \lim\limits_{n\to\infty} S_n(a, b) \equiv \lim\limits_{n\to\infty} \sum_{k=1}^{n} f(a + k\Delta x)\Delta x$.

Example

Let's apply the Riemann sum formula in the case of the 12-rectangle approximation to the function $f(x) = x^3 - 5x^2 + x + 10$ illustrated in Figure 5.9. The width of each rectangle is $\Delta x = \frac{4-(-1)}{12} = \frac{5}{12}$. The location of the right endpoint of the k^{th} rectangle is given by the formula

$$x_k = -1 + k\Delta x = -1 + k\frac{5}{12}\,.$$

The area of the k^{th} rectangle is equal to the height of the function $f(x_k)$ times the width Δx:

$$s_k = f(x_k)\Delta x = (x_k^3 - 5x_k^2 + x_k + 10)\frac{5}{12}.$$

The value for the 12-rectangle approximation to the area under the curve is

$$S_{12}(a,b) = \sum_{k=1}^{12} f(a + k\Delta x)\Delta x = \sum_{k=1}^{12}(x_k^3 - 5x_k^2 + x_k + 10)\frac{5}{12}.$$

Relax, we won't be doing the calculation by hand! We'll get the computer to calculate this summation for us. Go to live.sympy.org and type in the following expressions:

```
>>> n = 12.0;   xk = -1 + k*5/n;
>>> sk = (xk**3-5*xk**2+xk+10)*(5/n);
>>> summation( sk, (k,1,n) )
     11.802662...                        # the value of S_12(a,b)
```

The actual value of the area under the curve is given by

$$A(-1,4) = 12.91666\ldots.$$

Comparing our approximation $S_{12}(-1,4)$ with the true value $A(-1,4)$ we see that it is not very accurate. At least we got the first digit right! Let's see if we can do better.

More is better

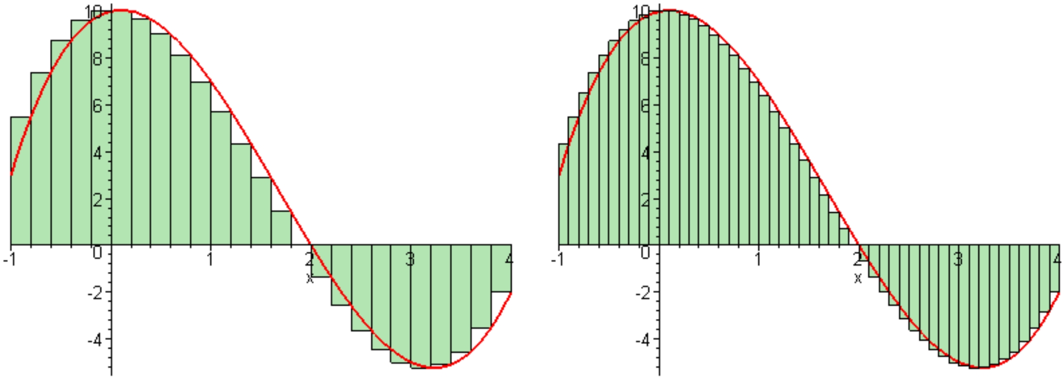
The 12-rectangle approximation is very low fidelity. You can *clearly* see some rectangles lie outside of the curve (overestimates), and some are too far inside the curve (underestimates). You might be wondering why we are wasting so much time to achieve such a lousy approximation. We have not been wasting our time. You see, the Riemann sum formula $S_n(a,b)$ gets better and better as you cut the region into smaller and smaller rectangles.

Using $n = 25$ rectangles, we obtain a better approximation:

$$S_{25}(a,b) = \sum_{k=1}^{25} f(a + k\Delta x)\Delta x = 12.4.$$

For $n = 50$, we obtain an even closer approximation:

$$S_{50}(a,b) = \sum_{k=1}^{50} f(a + k\Delta x)\Delta x = 12.6625.$$



(a) $n = 25$ (b) $n = 50$

Figure 5.10: An approximation to the area under the graph of the function $f(x) = x^3 - 5x^2 + x + 10$ using $n = 25$ and $n = 50$ rectangles.

For $n = 100$, the sum of the rectangles' areas starts to look pretttty much like the function. The calculation gives us $S_{100}(a, b) = 12.7906$.

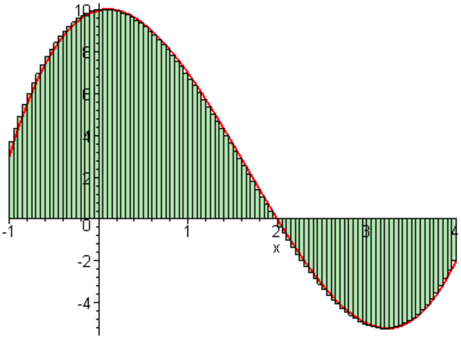


Figure 5.11: An approximation of the area under the function $f(x) = x^3 - 5x^2 + x + 10$ between $x = -1$ and $x = 4$ using $n = 100$ rectangles.

Using $n = 1000$ rectangles, we obtain an approximation to the area $S_{1000}(-1, 4) = 12.9041562$, which is accurate to the first decimal.

In the long run, when n grows really large, the Riemann sum approximations will get better and better and approach the true value of the area under the curve. Imagine cutting the region into $n = 10000$ rectangles; isn't $S_{10000}(-1, 4)$ a pretty accurate approximation of the actual area $A(-1, 4)$?

The integral as a limit

In the limit, as the number of rectangles n approaches ∞, the Riemann sum approximation to the area under the curve will become *arbitrarily close* to the true area:

$$\lim_{n\to\infty} \sum_{k=1}^{n} f(a + k\Delta x)\Delta x = A(a, b).$$

The definite integral between $x = a$ and $x = b$ is *defined* as the limit of a Riemann sum as n goes to infinity:

$$\int_a^b f(x)\,dx \equiv \lim_{n\to\infty} \sum_{k=1}^{n} f(a+k\Delta x)\Delta x.$$

Perhaps now the weird notation we use for integrals will start to make more sense to you. An integral is, literally, the sum of the function at the different sample points! In the limit as $n \to \infty$, the summation sign $\sum$ becomes an integral sign $\int$, and the step size Δx becomes an infinitesimally small step dx.

It is not computationally practical to make $n \to \infty$; we can simply stop at some finite n which produces the desired accuracy of approximation. The approximation using 1 million rectangles is accurate to the fourth decimal place, which you can verify by entering the following commands on `live.sympy.org`:

```
>>> n=1000000.0;
>>> xk=-1+k*5/n;
>>> sk=(xk**3-5*xk**2+xk+10)*(5/n);
>>> summation( sk, (k,1,n) )
     12.9166541666563
>>> integrate( x**3-5*x**2+x+10, (x,-1,4) ).evalf()
     12.9166666666667
```

Formal definition of the integral

We rarely compute integrals with Riemann sums. The Riemann sum is a theoretical construct, like the rise-over-run calculation that we use to define the derivative operation:

$$f'(x) = \lim_{\delta\to 0} \frac{f(x+\delta)-f(x)}{\delta}\,.$$

The integral is defined as the approximation of the area under the curve with infinitely many rectangles:

$$\int_a^b f(x)\,dx \equiv \lim_{n\to\infty} \sum_{k=1}^{n} f(a+k\Delta x)\Delta x, \quad \Delta x = \frac{b-a}{n}\,.$$

It is usually much easier to refer to a table of derivative formulas (see page 301) rather than compute a derivative starting from the formal definition and taking the limit $\delta \to 0$. Similarly, it is easier to refer to a table of integral formulas (also see page 301), rather than computing an integral by taking the limit as $n \to \infty$ of a Riemann sum.

Now that we have established a formal definition of the integral, we'll be able to understand why integral formulas are equivalent to derivative formulas applied in the opposite direction. In the next section we'll give a formal proof of the inverse relationship between the derivative operation and the integral operation.

Links

[Riemann sum wizard]
http://mathworld.wolfram.com/RiemannSum.html

5.14 The fundamental theorem of calculus

In Section 5.12 we defined the integral function $A_0(x)$ which corresponds to the calculation of the area under $f(x)$ starting from $x = 0$:

$$A_0(x) \equiv \int_0^x f(u)\, du.$$

We also discussed the notion of an antiderivative function: the function $F(x)$ is an antiderivative of $f(x)$ if $F'(x) = f(x)$.

A priori, these is no reason to suspect the integral function would be related to the derivative operation. The integral corresponds to the computation of an area, whereas the derivative operation computes the slope of a function. The fundamental theorem of calculus describes the relationship between derivatives and integrals.

Theorem (fundamental theorem of calculus) *Let $f(x)$ be a continuous function on the interval $[a,b]$, and let $c \in \mathbb{R}$ be a constant. Define the function $A_c(x)$ as follows:*

$$A_c(x) \equiv \int_c^x f(u)\, du.$$

Then, the derivative of $A_c(x)$ is equal to $f(x)$:

$$\frac{d}{dx}[A_c(x)] = f(x),$$

for any $x \in (a,b)$.

The fundamental theorem of calculus establishes an equivalence between the set of integral functions and the set of antiderivative functions:

$$A_c(x) = F(x) + C.$$

All integral functions $A_c(x)$ are antiderivatives of $f(x)$.

Differential calculus and integral calculus are two sides of the same coin. If you understand why the theorem is true, you will understand something very deep about calculus. Differentiation is the inverse operation of integration. Given a function $G(x) = \int g(x)dx$, we can obtain the function $g(x)$ by taking the derivative of $G(x)$: $G'(x) = g(x)$. The inverse relationship works the other way too. If you are given the derivative $h'(x)$ of some unknown function $h(x)$, you can find the function $h(x)$ (up to a constant), using integration: $h(x) + C = \int h'(x)dx$.

Got proof?

There is an unspoken rule in mathematics: when the word *theorem* appears in writing, it must be followed by the word *proof*. We therefore need to look into the proof of the fundamental theorem of calculus (FTC). It is not so important you understand the details of the proof, but I still recommend you read this subsection for your general math knowledge. If you are in a rush though, feel free to skip ahead.

Before we get to the proof of the FTC, we'll first introduce the *squeezing principle*, which we'll use in the proof. Suppose you have three functions, f, ℓ, and u, such that

$$\ell(x) \leq f(x) \leq u(x) \qquad \text{for all } x.$$

We say $\ell(x)$ is a *lower bound* on $f(x)$ since its graph is always below that of $f(x)$. Similarly, $u(x)$ is an *upper bound* on $f(x)$. We know the value of $f(x)$ is between $\ell(x)$ and $u(x)$.

Suppose $u(x)$ and $\ell(x)$ both converge to the same limit L:

$$\lim_{x \to a} \ell(x) = L, \quad \text{and} \quad \lim_{x \to a} u(x) = L.$$

Then it must be true that $f(x)$ also converges to the same limit:

$$\lim_{x \to a} f(x) = L.$$

This is true because the function f is *squeezed* between ℓ and u; it has no other choice than to converge to the same limit.

Proof of the fundamental theorem of calculus

For the sake of concreteness, let's use a fixed lower limit of integration $c = 0$. Our starting point is the graph of the function $f(x)$ and the definition of the integral function

$$A_0(x) \equiv \int_0^x f(u)\, du.$$

Our goal is to show that the derivative of the function $A_0(x)$ with respect to x is the function $f(x)$.

Recall the definition of the derivative $g'(x) = \lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x)}{\epsilon}$. If we want to find the derivative of $A_0(x)$, we must compute the difference $A_0(x + \epsilon) - A_0(x)$

and then divide by ϵ. Using the definition of the integral function $A_0(x)$, we obtain

$$\begin{aligned} A_0(x+\epsilon) - A_0(x) &= \int_0^{x+\epsilon} f(t)\,dt - \int_0^{x} f(t)\,dt \\ &= \int_x^{x+\epsilon} f(t)\,dt. \end{aligned}$$

Figure 5.12 illustrates the region corresponding to this difference $A_0(x+\epsilon)-A_0(x)$. The region is a long, vertical strip of width ϵ, and a height that varies according to $f(x)$:

$$\int_x^{x+\epsilon} f(t)\,dt \approx \underbrace{\text{width}}_{\epsilon} \times \underbrace{\text{height}}_{?}.$$

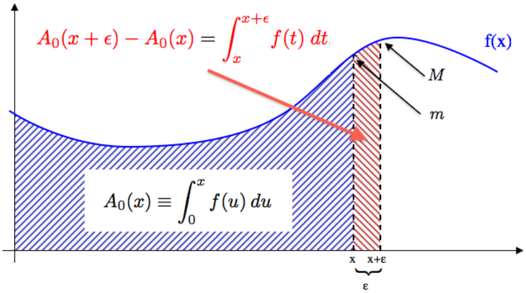

$A_0(x+\epsilon) - A_0(x) = \int_x^{x+\epsilon} f(t)\,dt$
f(x)
M
m
$A_0(x) \equiv \int_0^x f(u)\,du$
x
x+ε
ε

Figure 5.12: The difference $A_0(x+\epsilon) - A_0(x)$ corresponds to the integral of $f(x)$ between x and $x+\epsilon$.

Define the numbers M and m that correspond to the maximum and minimum values of the function $f(x)$ on the interval $[x, x+\epsilon]$:

$$M \equiv \max_{t\in[x,x+\epsilon]} f(t) \qquad \text{and} \qquad m \equiv \min_{t\in[x,x+\epsilon]} f(t).$$

By definition, the quantities ϵm and ϵM provide a lower and an upper bound on the quantity we're trying to study:

$$\epsilon m \le \int_x^{x+\epsilon} f(t)\,dt \le \epsilon M.$$

Recall from the theorem statement that f is *continuous*. If f is continuous, then as $\epsilon \to 0$, we'll have

$$\lim_{\epsilon \to 0} f(x+\epsilon) = f(x).$$

In fact, as $\epsilon \to 0$, all the values of f on the shortening interval $[x, x+\epsilon]$ will approach $f(x)$. In particular, both the minimum value m and the maximum value M will approach $f(x)$:

$$\lim_{\epsilon \to 0} f(x+\epsilon) = f(x) = \lim_{\epsilon \to 0} m = \lim_{\epsilon \to 0} M.$$

So, starting from the inequality

$$\epsilon m \leq \int_x^{x+\epsilon} f(t)\ dt \leq \epsilon M,$$

and taking the limit as $\epsilon \to 0$, we obtain

$$\begin{aligned} \lim_{\epsilon \to 0} \epsilon m &\leq \lim_{\epsilon \to 0} \int_x^{x+\epsilon} f(t)\ dt \leq \lim_{\epsilon \to 0} \epsilon M, \\ \lim_{\epsilon \to 0} \epsilon f(x) &\leq \lim_{\epsilon \to 0} \int_x^{x+\epsilon} f(t)\ dt \leq \lim_{\epsilon \to 0} \epsilon f(x). \end{aligned}$$

Then applying the squeezing principle, we obtain

$$\lim_{\epsilon \to 0} \int_x^{x+\epsilon} f(t)\ dt = \lim_{\epsilon \to 0} \epsilon f(x). \tag{†}$$

Let's see how this expression fits into the formula for the derivative of $A_0(x)$:

$$\begin{aligned} A_0'(x) &= \lim_{\epsilon \to 0} \frac{A_0(x+\epsilon) - A_0(x)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\int_x^{x+\epsilon} f(t)\ dt}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\epsilon f(t)}{\epsilon} && \text{(by using equation (†))} \\ &= f(x) \lim_{\epsilon \to 0} \frac{\epsilon}{\epsilon} \\ &= f(x). \end{aligned}$$

Thus we have proved that $A'_0(x) = f(x)$. The argument presented did not depend on the choice $c = 0$ we made in the beginning, so the statement $A'_c(x) = f(x)$ is true for all $c \in \mathbb{R}$. □

Applications

You can use the fundamental theorem of calculus to check your answers to indefinite integral questions.

Example: Integral verification Suppose a friend tells you that

$$\int \ln(x)dx = x\ln(x) - x + C,$$

but he's a shady character and you don't trust him. How can you check his answer? If you had a smartphone handy, you could check the answer on `live.sympy.org`. What if you just have some old-school pen and paper? If $x\ln(x) - x$ is really the integral of $\ln(x)$, then by the fundamental theorem of calculus, we should obtain $\ln(x)$ if we take the derivative. Let's check:

$$\begin{aligned}\frac{d}{dx}[x\ln(x) - x] &= \overbrace{\frac{d}{dx}[x]\ln(x) + x\left[\frac{d}{dx}\ln(x)\right]}^{\text{product rule}} - \frac{d}{dx}[x] \\ &= 1\ln(x) + x\frac{1}{x} - 1 = \ln(x).\end{aligned}$$

This time, your shady friend is correct.

Discussion

Integration and differentiation are inverse operations

You previously studied the inverse relationship for functions. Recall that for any *bijective* function f (a one-to-one relationship) there exists an *inverse function* f^{-1}

that *undoes* the effects of f:

$$(f^{-1} \circ f)(x) \equiv f^{-1}(f(x)) = 1x$$

and also

$$(f \circ f^{-1})(y) \equiv f(f^{-1}(y)) = 1y.$$

The integral is the "inverse operation" of the derivative. If you perform the integral operation followed by the derivative operation on some function, you'll obtain the same function:

$$\left(\frac{d}{dx} \circ \int dx\right) f(x) \equiv \frac{d}{dx} \int_c^x f(u)\, du = f(x).$$

Note we need a new variable u inside the integral since x is already used to denote the upper limit of integration.

Alternately, if you compute the derivative followed by the integral, you will obtain the original function $f(x)$ (up to a constant):

$$\left(\int dx \circ \frac{d}{dx}\right) f(x) \equiv \int_c^x f'(u)\, du = f(x) + C.$$

What next?

If integration is nothing more than backward differentiation, and if you already know differentiation inside out from differential calculus, you might be wondering what you are going to do during an entire semester of integral calculus. For all intents and purposes, if you understand the conceptual material in this section, you understand integral calculus. Give yourself a pat on the back—you are done.

The Establishment, however, not only wants you to know the concepts of integral calculus; you must also become proficient in computing integrals of functions. Thus, you'll need to practice the techniques of integration presented in the next section. There are a bunch of techniques that allow you to integrate complicated functions. For example, if I asked you to integrate $f(x) = \sin^2(x) = (\sin(x))^2$

from 0 to π and you look at the formula sheet, you won't find a function $F(x)$ with a derivative that equals $f(x)$. So how do we solve

$$\int_0^\pi \sin^2(x)\ dx = \ ?$$

One way to approach this problem is with the double-angle trigonometric identity, which states that $\sin^2(x) = \frac{1-\cos(2x)}{2}$. Using this identity, we can proceed as follows:

$$\int_0^\pi \sin^2(x) dx = \int_0^\pi \left[\frac{1}{2} - \frac{1}{2}\cos(2x)\right] dx = \underbrace{\frac{1}{2}\int_0^\pi 1\ dx}_{T_1} - \underbrace{\frac{1}{2}\int_0^\pi \cos(2x)\ dx}_{T_2}.$$

We are allowed to split the integral into two parts and take the constants $\frac{1}{2}$ out of the integrals because integration is a *linear* operation.

The integral we want to calculate can be computed as the difference of two terms:

$$\int_0^\pi \sin^2(x)\, dx \quad = \quad T_1 \ - \ T_2.$$

Let's compute the terms one by one.

The value of the integral in the first term is

$$T_1 = \frac{1}{2}\int_0^\pi 1\ dx = \frac{1}{2}x\bigg|_0^\pi = \frac{\pi - 0}{2} = \frac{\pi}{2}.$$

The value of the second term is

$$T_2 = \frac{1}{2}\int_0^\pi \cos(2x)\ dx = \frac{1}{4}\sin(2x)\bigg|_0^\pi = \frac{\sin(2\pi) - \sin(0)}{4} = \frac{0-0}{4} = 0.$$

We find the final answer for the integral to be

$$\int_0^\pi \sin^2(x)\ dx = T_1 - T_2 = \frac{\pi}{2} - 0 = \frac{\pi}{2}.$$

Do you see how integration can quickly become tricky? You need to learn all kinds of tricks to solve integrals. I can teach you all the necessary tricks, but to become proficient you can't solely read—you need to *practice* the techniques. Promise me you will practice! As my student, I expect nothing less than total ass-kicking of the questions you'll face on the final exam.

5.15 Techniques of integration

The operation of "taking the integral" of some function is usually much more complicated than that of taking the derivative. You can take the derivative of *any* function—no matter how complex—by using the product rule, the chain rule, and the derivative formulas. This is not true for integrals.

Plenty of integrals have no *closed-form solution*, meaning the function has no antiderivative. There is no simple procedure to follow such that you input a function and "turn the crank" until the integral comes out. Integration is a bit of an art.

Which functions *can* we integrate, and how? Back in the day, scientists collected big tables with integral formulas for various complicated functions. We can use these tables to *look up* a specific integral formula. Such table is given on page 462 in the back of the book.

We can also learn some *integration techniques* to help make complicated integrals simpler. Think of the techniques presented in this section as *adapters*. You can reach for these adapters when the function you need to integrate doesn't appear in your table of integrals, but a *similar* one is found in the table.

A note to all our students in the audience who are taking an integral calculus course. These integration techniques are exactly the skills you'll be expected to demonstrate on the final. Instead of using the table of integrals to look up complicated integrals, you'll need to know how to make your own table.

For people interested in learning physics, I'll honestly tell you that if you skip this next section you won't miss much. You should read the important section on *substitution*, but there's no need to read the details of all the recipes for integrating things. For most intents and purposes, once you understand what an integral is, you can use a computer to calculate it. A good tool for calculating integrals is the computer algebra system at `live.sympy.org`.

```
>>> integrate( sin(x) )
     -cos(x)
>>> integrate( x**2*exp(x) )
     x**2*exp(x) - 2*x*exp(x) + 2*exp(x)
```

You can use `sympy` for all your integration needs.

A comment to those of you reading this book for *general culture*, without the added stress of homework and exams. Consider the next dozen pages as a snapshot of the daily life of the undergraduate experience in science. Try to visualize the life of first-year science students, busy integrating things they don't want to integrate for many, long hours. Picture some unlucky science student locked in her room, crunching calculus while hundreds of dangling integrals scream for attention, keeping her from hanging with friends.

Actually, it is not that bad. There are, like, four tricks to learn. If you **practice**, you can learn all of them in a week or so. Mastering these four tricks is essentially the purpose of the entire integral calculus class. If you understand the material in this section, you'll be done with integral calculus and you'll have two months to chill.

Substitution

Say you're integrating some complicated function that contains a square root $\sqrt{x}$. You wonder how to compute this integral:

$$\int \frac{1}{x-\sqrt{x}}\, dx \ = ?$$

Sometimes you can simplify an integral by *substituting* a new variable into the expression. Let $u = \sqrt{x}$. Substitution is like search-and-replace in a word processor. Every time you see the expression $\sqrt{x}$, replace it with u:

$$\int \frac{1}{x-\sqrt{x}}\, dx = \int \frac{1}{u^2-u}\, dx.$$

Note we also replaced $x = (\sqrt{x})^2$ with u^2.

We're not done yet. To change from the x variable to the u variable, we must also change dx to du. Can we simply replace dx with du? Unfortunately no, otherwise it would be like saying the "short step" du is equal in length to the "short step" dx, which is only true for the trivial substitution $u = x$.

To find the relation between the infinitesimals, we take the derivative:

$$u(x) = \sqrt{x} \quad \Rightarrow \quad u'(x) = \frac{du}{dx} = \frac{1}{2\sqrt{x}}.$$

For the next step, I need you to stop thinking about the expression $\frac{du}{dx}$ as a whole, and instead think about it as a rise-over-run fraction that can be split. Let's move the *run* dx to the other side of the equation:

$$du = \frac{1}{2\sqrt{x}}\, dx.$$

Next, to isolate dx, multiply both sides by $2\sqrt{x}$:

$$dx = 2\sqrt{x}\, du = 2u\, du,$$

where we use the fact that $u = \sqrt{x}$ in the last step.

We now have an expression for dx expressed entirely in terms of the variable u. After the substitution, the integral looks like

$$\int \frac{1}{x - \sqrt{x}}\, dx = \int \frac{1}{u^2 - u} 2u\, du = \int \frac{2}{u - 1}\, du.$$

We can recognize the general form of the function inside the integral, $f(u) = \frac{2}{u-1}$, to be similar to the function $f(u) = \frac{1}{u}$. Recall that the integral of $\frac{1}{u}$ is $\ln(u)$. Accounting for the -1 horizontal shift and the factor of 2 in the numerator, we obtain the answer:

$$\int \frac{1}{x - \sqrt{x}}\, dx = \int \frac{2}{u - 1}\, du = 2\ln(u - 1) = 2\ln(\sqrt{x} - 1).$$

Note in the last step, we changed back to the x variable to give the final answer. The variable u exists only in our calculation. We invented it out of thin air when we said, "Let $u = \sqrt{x}$" in the beginning.

Thanks to the substitution, the integral becomes simpler since we eliminate the square roots. The extra u that comes from the expression $dx = 2u\,du$ cancels with one of the us in the denominator, making things even simpler. In practice, substituting inside f is the easy part. The hard part is making sure our choice of substitution leads to a replacement for dx that helps to simplify the integral.

For definite integrals—that is, integrals with limits of integration—there is an extra step we need to take when changing variables: we must change the x limits of integration to u limits. In our expression, when changing to the u variable, we write

$$\int_a^b \frac{1}{x - \sqrt{x}}\,dx = \int_{u(a)}^{u(b)} \frac{2}{u-1}\,du.$$

Say we are asked to compute the definite integral between $x = 4$ and $x = 9$ for the same expression. In this case, the new limits are $u = \sqrt{4} = 2$ and $u = \sqrt{9} = 3$, and we have

$$\int_4^9 \frac{1}{x - \sqrt{x}}\,dx = \int_2^3 \frac{2}{u-1}\,du = 2\ln(u-1)\bigg|_2^3 = 2(\ln(2) - \ln(1)) = 2\ln(2).$$

Let's recap. Substitution involves three steps:

1. Replace all occurrences of $u(x)$ with u.
2. Replace dx with $\frac{1}{u'(x)}du$.
3. If there are limits, replace the x limits with u limits.

If the resulting integral is simpler to solve, then good for you!

Example Find $\int \tan(x)\ dx$. We know $\tan(x) = \frac{\sin(x)}{\cos(x)}$, so we can use the substitution $u = \cos(x)$, $du = -\sin(x)dx$ as follows:

$$\begin{aligned}
\int \tan(x)dx &= \int \frac{\sin(x)}{\cos(x)}dx \\
&= \int \frac{-1}{u}du \\
&= -\ln|u| + C \\
&= -\ln|\cos(x)| + C.
\end{aligned}$$

Integrals of trig functions

Because $\sin$, $\cos$, $\tan$, and the other trig functions are related, we can often express one function in terms of another in order to simplify integrals.

Recall the trigonometric identity,

$$\cos^2(x) + \sin^2(x) = 1,$$

which is the statement of Pythagoras' theorem.

If we choose to make the substitution $u = \sin(x)$, we can replace all kinds of trigonometric terms with the new variable u:

$$\begin{aligned}
\sin^2(x) &= u^2, \\
\cos^2(x) &= 1 - \sin^2(x) = 1 - u^2, \\
\tan^2(x) &= \frac{\sin^2(x)}{\cos^2(x)} = \frac{u^2}{1-u^2}.
\end{aligned}$$

Of course the change of variable $u = \sin(x)$ means you must also change the $du = u'(x)dx = \cos(x)dx$, so there better be something to cancel this $\cos(x)$ term in the integral.

Let me show you one example where things work perfectly. Suppose m is some arbitrary number and you need to integrate:

$$\int (\sin(x))^m \cos^3(x)\ dx \equiv \int \sin^m(x) \cos^3(x)\ dx.$$

This integral contains m powers of the $\sin$ function and three powers of the $\cos$ function. Let us split the $\cos$ term into two parts:

$$\int \sin^m(x)\cos^3(x)\;dx = \int \sin^m(x)\cos^2(x)\cos(x)\;dx.$$

Making the change of variables ($u = \sin(x)$ and $du = \cos(x)dx$) means we can replace $\sin^m(x)$ by u^m, and $\cos^2(x) = 1 - u^2$ in the above expression to obtain

$$\int \sin^m(x)\cos^2(x)\cos(x)\;dx = \int u^m\left(1-u^2\right)\cos(x)\;dx.$$

Conveniently, we happen to have $dx = \frac{1}{\cos(x)}du$, so the complete change-of-variable step is

$$\int \sin^m(x)\cos^2(x)\cos(x)\;dx = \int u^m\left(1-u^2\right)\;du.$$

This is what I was talking about earlier when I mentioned "having an extra $\cos(x)$" to cancel the one that appears as a result of the $dx \to du$ change.

What is the answer then? It is a simple integral of a polynomial:

$$\begin{aligned}\int u^m\left(1-u^2\right)\;du &= \int \left(u^m - u^{m+2}\right)\;du \\ &= \frac{1}{m+1}u^{m+1} - \frac{1}{m+3}u^{m+3} \\ &= \frac{1}{m+1}\sin^{m+1}(x) - \frac{1}{m+3}\sin^{m+3}(x).\end{aligned}$$

You might be wondering how useful this substitution technique actually is. I mean, how often will you need to integrate such particular combinations of $\sin$ and $\cos$ powers, where substitution works perfectly? You might be surprised! Sin and cos functions are used often in this thing called the *Fourier transform*, which is a way of expressing a sound wave $f(t)$ in terms of the frequencies it contains.

Also, integrals with trig functions are known favourites of teachers to ask on exams. A trig substitution question will test if you can perform substitutions, *and* teachers use them to check whether you remember all the trigonometric identities, which you are supposed to have learned in high school.

Are there other trig substitution tricks you should know? On an exam, you should try any possible substitution you can think of, combined with any trigonometric identity that seems to simplify things. Some common substitutions are described below.

Cos

Just as we can substitute $\sin$, we can also substitute $u = \cos(x)$ and use $\sin^2(x) = 1 - u^2$. Again, this substitution only makes sense when there is a leftover $\sin$ somewhere in the integral that can cancel with the $\sin$ in $dx = \frac{-1}{\sin x}\,du$.

Tan and sec

We can get some more mileage out of the trigonometric identity $\cos^2(x)+\sin^2(x) = 1$. Dividing both sides of this identity by $\cos^2(x)$ gives us

$$1 + \tan^2(x) = \sec^2(x) \equiv \frac{1}{\cos^2(x)}.$$

This is useful since $u = \tan(x)$ gives $du = \sec^2(x)dx$, allowing us to "kill" even powers of $\sec^2(x)$ in integrals of the form

$$\int \tan^m(x)\sec^n(x)\,dx.$$

Even powers of sin and cos

There are other trigonometric identities, called half-angle and double-angle formulas, which give us formulas such as

$$\sin^2(x) = \frac{1}{2}(1 - \cos(2x)), \qquad \cos^2(x) = \frac{1}{2}(1 + \cos(2x)).$$

These identities are useful if you need to integrate even powers of $\sin$ and $\cos$.

Example How can we find $I = \int \sin^2(x)\cos^4(x)\,dx$? Let's find out:

$$
\begin{aligned}
I &= \int \sin^2(x)\cos^4(x)\;dx \\
&= \int \left(\frac{1}{2}(1-\cos(2x))\right)\left(\frac{1}{2}(1+\cos(2x))\right)^2 dx \\
&= \frac{1}{8}\int \big(1-\cos^2(2x)+\cos(2x)-\cos^3(2x)\big)\,dx \\
&= \frac{1}{8}\int \big(1-\cos^2(2x)+\cos(2x)-\cos^2(2x)\cos(2x)\big)\,dx \\
&= \frac{1}{8}\int \left(1-\frac{1}{2}(1+\cos(4x))+\cancel{\cos(2x)}-(\cancel{1}-\sin^2(2x))\underline{\cos(2x)}\right)dx \\
&= \frac{1}{8}\int \left(\frac{1}{2}-\frac{1}{2}\cos(4x)+\underbrace{\sin^2(2x)}_{u^2}\cos(2x)\right)dx \qquad (\text{let } u=\sin(2x)) \\
&= \frac{1}{8}\left(\frac{x}{2}-\frac{\sin(4x)}{8}+\frac{\sin^3(2x)}{6}\right) \\
&= \frac{x}{16}-\frac{\sin(4x)}{64}+\frac{\sin^3(2x)}{48}+C.
\end{aligned}
$$

There is no limit to the number of combinations of simplification steps you can try. On a homework question or an exam, the teacher will ask for something simple. Your job is to find the correct substitution.

Sneaky example Sometimes, the required substitution is not obvious at all, as in the case of $\int \sec(x)dx$. To find the integral, you need the following trick: multiply and divide by $\tan(x)+\sec(x)$:

$$
\begin{aligned}
\int \sec(x)\,dx &= \int \sec(x)\;1\,dx \\
&= \int \sec(x)\frac{\tan(x)+\sec(x)}{\tan(x)+\sec(x)}\;dx
\end{aligned}
$$

$$
\begin{aligned}
&= \int \frac{\sec^2(x) + \sec(x)\tan(x)}{\tan(x) + \sec(x)}\, dx \\
&= \int \frac{1}{u} du \\
&= \ln|u| + C = \ln|\tan(x) + \sec(x)| + C,
\end{aligned}
$$

where, in the fourth line of solving, we used the substitution $u = \tan(x) + \sec(x)$ and $du = (\sec^2(x) + \tan(x)\sec(x))dx$.

I highly recommend you practice all the examples you can get your hands on. Don't bother memorizing any recipes though; you can do just as well with trial and error.

Trig substitution

Often when calculating integrals for physics, we run into terms of the form $\sqrt{a^2 - x^2}$, $\sqrt{a^2 + x^2}$, or $\sqrt{x^2 - a^2}$, which can be difficult to handle directly. In each of these three instances, we can perform a *trig substitution*, replacing x with one of the trigonometric functions $a\sin(\theta)$, $a\tan(\theta)$, or $a\sec(\theta)$ to obtain a simpler integral.

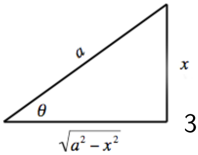
Sine substitution

Consider an integral that contains an expression of the form $\sqrt{a^2 - x^2}$. By applying the substitution $x = a\sin\theta$, the complicated square-root expression is simplified:

$$\sqrt{a^2 - x^2} = \sqrt{a^2 - a^2\sin^2\theta} = a\sqrt{1 - \sin^2\theta} = a\cos\theta.$$

The simplification is possible because of the identity $\cos^2\theta = 1 - \sin^2\theta$. The transformed integral now involves a trigonometric function that we know how to integrate.

Once we find the integral in terms of θ, we need to look at the answer and convert the various θ expressions therein



back to the original variables x and a:

$$\sin\theta = \frac{x}{a},\ \cos\theta = \frac{\sqrt{a^2-x^2}}{a},\ \tan\theta = \frac{x}{\sqrt{a^2-x^2}},$$

$$\csc\theta = \frac{a}{x},\quad \sec\theta = \frac{a}{\sqrt{a^2-x^2}},\quad \cot\theta = \frac{\sqrt{a^2-x^2}}{x}.$$

Example 1 Calculate $\int \sqrt{1-x^2}\,dx$.

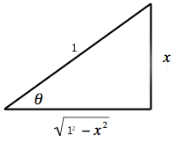
We can approach the problem by making the $\sin$ substitution with $a = 1$:

$$x = \sin\theta, \qquad dx = \cos\theta\ d\theta.$$

We proceed:

$$\begin{aligned}
\int \sqrt{1-x^2}\ dx &= \int \sqrt{1-\sin^2\theta}\cos\theta\ d\theta \\
&= \int \cos^2\theta\ d\theta \\
&= \frac{1}{2}\int [1+\cos 2\theta]\ d\theta \\
&= \frac{1}{2}\theta + \frac{1}{4}\sin 2\theta \\
&= \frac{1}{2}\theta + \frac{1}{2}\sin\theta\cos\theta \\
&= \frac{1}{2}\sin^{-1}(x) + \frac{1}{2}\frac{x}{1}\frac{\sqrt{1-x^2}}{1}.
\end{aligned}$$

In the last step, we use the triangle diagram to "read" the values of θ, $\sin\theta$, and $\cos\theta$ from the triangle. The substitution $x = \sin\theta$ means the hypotenuse in the diagram is of length 1, and the opposite side is of length x.

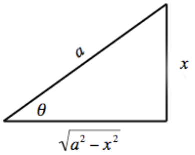


Example 2 Compute $\int \sqrt{\frac{a+x}{a-x}}\, dx$.

We can rewrite this fraction as

$$\sqrt{\frac{a+x}{a-x}} = \sqrt{\frac{a+x}{a-x}\frac{1}{1}} = \sqrt{\frac{a+x}{a-x}\frac{a+x}{a+x}} = \frac{a+x}{\sqrt{a^2-x^2}}.$$

We make the substitution $x = a\sin\theta$, and $dx = a\cos\theta d\theta$ and proceed as follows:



$$\begin{aligned}\int \frac{a+x}{\sqrt{a^2-x^2}}dx &= \int \frac{a+a\sin\theta}{\cancel{a\cos\theta}}\cancel{a\cos\theta}\, d\theta \\ &= a\int [1+\sin\theta]\, d\theta \\ &= a\left[\theta - \cos\theta\right] \\ &= a\sin^{-1}\left(\frac{x}{a}\right) - a\frac{\sqrt{a^2-x^2}}{a} \\ &= a\sin^{-1}\left(\frac{x}{a}\right) - \sqrt{a^2-x^2}.\end{aligned}$$

Tan substitution

When an integral contains $\sqrt{a^2+x^2}$, use the substitution,

$$x = a\tan\theta, \qquad dx = a\sec^2\theta d\theta.$$

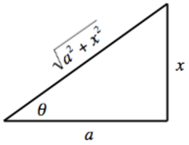
Because of the identity $1+\tan^2\theta = \sec^2\theta$, the square root expression will simplify to

$$\sqrt{a^2+x^2} = \sqrt{a^2+a^2\tan^2\theta} = a\sqrt{1+\tan^2\theta} = a\sec\theta.$$

Simplification is a good thing. Simplification makes it much easier to find the integral in terms of θ than in terms of $\sqrt{a^2+x^2}$.

Once you calculate the integral in terms of θ, you can convert the answer back into x coordinates. To do this, you need to draw a triangle labelled according to your substitution:

$$\tan\theta = \frac{x}{a} = \frac{\textsf{opp}}{\textsf{adj}}.$$



The equivalent of $\sin\theta$ in terms of x is $\sin\theta \equiv \frac{\textsf{opp}}{\textsf{hyp}} = \frac{x}{\sqrt{a^2+x^2}}$. Similarly, the other trigonometric functions are defined as various ratios of a, x, and $\sqrt{a^2+x^2}$.

Example Calculate $\int \frac{1}{x^2+1}\,dx$.

The denominator of this function is equal to $\left(\sqrt{1+x^2}\right)^2$. The form $1+x^2$ suggests we can probably substitute $x = \tan\theta$, then use the identity $1+\tan^2\theta = \sec^2\theta$. Testing this substitution, we obtain $dx = \sec^2\theta\,d\theta$. Thus,

$$\begin{aligned}\int \frac{1}{x^2+1}\,dx &= \int \frac{1}{\tan^2\theta+1}\sec^2\theta\,d\theta \\ &= \int \frac{1}{\sec^2\theta}\sec^2\theta\,d\theta \\ &= \int 1\;d\theta \\ &= \theta \\ &= \tan^{-1}(x) + C.\end{aligned}$$

Obfuscated example What if the denominator doesn't look like x^2+1? What if, instead, we have a general second-degree polynomial, such as

$$\frac{1}{y^2-6y+10}\,?$$

How do we integrate a this function? If there were no $-2y$ term, we'd be able to use the tan substitution. Or perhaps you could look up the formula $\int \frac{1}{x^2+1}dx =$

$\tan^{-1}(x)$ in the table of integrals. Alas, there is no formula to be found in the table for

$$\int \frac{1}{y^2 - 6y + 10}\, dy.$$

We'll need another route, and we'll start by following the good old substitution technique $u = \ldots$, along with a high school algebra trick called "completing the square." This route will help us rewrite the fraction inside the integral so the integral looks like $(y-h)^2 + k$ with no linear term.

First, find "by inspection" the values of h and k such that:

$$\frac{1}{y^2 - 6y + 10} = \frac{1}{(y-h)^2 + k} = \frac{1}{(y-3)^2 + 1}.$$

The "square completed" quadratic expression has no linear term. Now we'll use the substitution $x = y - 3$ and $dx = dy$ to obtain an integral we know how to solve:

$$\int \frac{1}{(y-3)^2 + 1}\, dy = \int \frac{1}{x^2 + 1}\, dx = \tan^{-1}(x) = \tan^{-1}(y-3).$$

Sec substitution

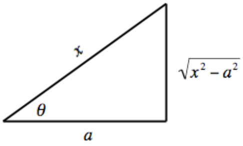
In the previous two sections, we learned how to handle $\sqrt{a^2 - x^2}$ and $\sqrt{x^2 + a^2}$. One more option remains: $\sqrt{x^2 - a^2}$, which we approach using the secant substitution.

Recall the trigonometric identity $1 + \tan^2\theta = \sec^2\theta$ or, rewritten differently,

$$\sec^2\theta - 1 = \tan^2\theta.$$

The appropriate substitution for terms like $\sqrt{x^2 - a^2}$ is

$$x = a\sec\theta, \qquad dx = a\tan\theta\sec\theta\, d\theta.$$



The substitution procedure is the same as in previous cases of the $\sin$ substitution and the $\tan$ substitution we discussed, so we won't elaborate here in detail. We can label the sides of the triangle accordingly, as

$$\sec\theta = \frac{x}{a} = \frac{\textsf{hyp}}{\textsf{adj}}.$$

We'll use this triangle when converting back from θ to x in the final steps of the integral calculation.

Interlude

By now, things are starting to get pretty tight for your calculus teacher. You are beginning to understand how to "handle" any kind of integral he can throw at you: polynomials, fractions with x^2, plus or minus a^2, and square roots. He can't even fool you with dirty trigonometric tricks involving $\sin$, $\cos$, and $\tan$, since you know about these, too. Are there any integrals left that he can drop on the exam to trick you up?

Substitution is the most important integration technique. Recall the steps involved: (1) the choice of substitution $u = \ldots$, (2) the associated dx to du change, and (3) the change in the limits of integration required for definite integrals. With medium to advanced substitution skills, you'll score at least an 80% on your integral calculus final.

What will the remaining 20% of the exam depend on? There are two more recipes to go. I know all these tricks that I've been throwing at you during the last ten pages may seem arduous and difficult to understand, but this is what you got yourself into when you signed up for the course "integral calculus." In this course, there are lots of *integrals* and you *calculate* them.

The good news is that we are almost done. Only one more "trick" remains, and afterward, I'll finally tell you about the *integration by parts* procedure, which is very useful.

Don't bother memorizing the steps in each problem. The correct substitution of $u = \ldots$ will be different in each problem. Think of integration techniques as general recipe guidelines you must adapt based on the ingredients available to you at the moment of cooking. You can always return to this section when faced with a complicated integral; check to see which of this section's examples looks the most similar and follow the same approach.

Partial fractions

Suppose you need to integrate a rational function $\frac{P(x)}{Q(x)}$, where P and Q are polynomials.

For example, you could be asked to integrate

$$\frac{P(x)}{Q(x)} = \frac{ax+b}{cx^2+dx+e},$$

where a, b, c, d, and e are arbitrary constants. To get even more specific, let's say you are asked to calculate

$$\int \frac{3x+1}{x^2+x}\,dx.$$

By magical powers, I can transform the function in this integral into two *simple fractions*:

$$\int \frac{3x+1}{x^2+x}\,dx \qquad = \qquad \int \frac{1}{x}\,dx \quad + \quad \int \frac{2}{x+1}\,dx.$$

We split the complicated-looking rational expression into two partial fractions.

Now that the hard part is done, all that remains is to compute the two integrals. Recall that $\frac{d}{dx}\ln(x) = \frac{1}{x}$, so the integrals will give $\ln$-like terms. The final answer is

$$\int \frac{3x+1}{x^2+x}\,dx = \ln|x| + 2\ln|x+1| + C.$$

How did I split the problem into partial fractions? Is it really magic or is there a method? The answer is, a little bit of both. My method was to *assume* the existence of constants A and B such that

$$\frac{3x+1}{x^2+x} = \frac{3x+1}{x(x+1)} = \frac{A}{x} + \frac{B}{x+1}.$$

Then I solved the above equation for A and B by computing the sum of the two fractions:

$$\frac{3x+1}{x(x+1)} = \frac{A(x+1)+Bx}{x(x+1)}.$$

The magic involves the fact that we can solve for *two* unknowns in *one* equation. The relevant part of the equation is the numerator, because both sides have the same denominator. To find A and B, we must solve

$$3x+1 = (3)x + (1)1 = A(x+1) + Bx = (A+B)x + (A)1.$$

We solve by grouping the unknown constants into two terms: a term involving x (the linear term) and a constant term. On the left-hand side of the equation, the constant term is equal to 1, and on the right-hand side the constant coefficient is A, so $A = 1$. Similarly, we can deduce that $B = 2$ from the equality of the coefficients of the linear terms $3 = A + B$, having found $A = 1$ in the first step.

Another way of finding the values of the unknowns A and B is by evaluating the numerator equation

$$3x+1 = A(x+1) + Bx$$

at different values of x. This equation must hold true for all values of the variable x. The input $x = 0$ gives us $1 = A$, and inputting $x = -1$ gives $-2 = -B$, so $B = 2$.

The above problem highlights the power of the *partial fractions* method for attacking integrals of polynomial fractions $\frac{P(x)}{Q(x)}$. Most of the work involves factoring the denominator and finding the unknowns. Then, once you've split the problem into partial fractions, some simple calculus steps finish the job. Some people call this method *separation of quotients*, but whatever you call it, it's clear that being able to split a fraction into multiple parts is very helpful:

$$\frac{3x+1}{x^2+x} = \frac{A}{x} + \frac{B}{x+1}.$$

How many parts will the fraction $\frac{P(x)}{Q(x)}$ split into? What will each part look like? The answer is that there will be as many parts as the degree of the polynomial

$Q(x)$, which is located in the fraction's denominator. Each part will consist of one of the factors of $Q(x)$.

1. Factor the denominator $Q(x)$ as a product of factors. For example, $Q(x) = x^3+4x^2+5x+2$ can be factored as $Q(x) = (x+1)^2(x+2)$. For each factor of $Q(x)$, *assume* an appropriate partial fraction term on the right-hand side of the equation. There are three types of factors to consider:
   - Simple factors, like $(x-\alpha)^1$. For each simple factor, you should *assume* a partial fraction of the form
     $$\frac{A}{x-\alpha}.$$
   - Repeated factors, like $(x-\beta)^n$, for which we assume the existence of n different terms on the right-hand side:
     $$\frac{B}{x-\beta}+\frac{C}{(x-\beta)^2}+\cdots+\frac{F}{(x-\beta)^n}.$$
   - If one of the factors is a polynomial ax^2+bx+c that cannot be factored, such as x^2+1, we must preserve this portion as a whole and assume that a term of the form
     $$\frac{Gx+H}{ax^2+bx+c}$$
     exists on the right-hand side.
2. Add all the parts on the equation's right-hand side by first cross-multiplying each part in order to bring all fractions to a common denominator, and then adding the fractions together. If you followed the steps correctly in Part 1, the *least common denominator* (LCD) will turn out to be $Q(x)$, and both sides will have the same denominator. Solve for the unknown coefficients $A, B, C, \ldots$ in the numerators. Find the coefficients of each power of x on the right-hand side and set them equal to the corresponding coefficient in the numerator $P(x)$ of the left-hand side.
3. Use the appropriate integral formula for each kind of term:

- For simple factors, use

$$\int \frac{1}{x-\alpha}\, dx = A\ln|x-\alpha| + C.$$

- For higher powers in the denominator, use

$$\int \frac{1}{(x-\beta)^m}\, dx = \frac{1-m}{(x-\beta)^{m-1}} + C.$$

- For the quadratic denominator terms with "matching" numerator terms, use

$$\int \frac{2ax+b}{ax^2+bx+c}\, dx = \ln|ax^2+bx+c| + C.$$

  For quadratic terms with only a constant in the numerator, use a two-step substitution process. First, change x to the complete-the-square variable $y = x - h$:

$$\int \frac{1}{ax^2+bx+c}\, dx = \int \frac{1/a}{(x-h)^2+k}\, dx = \frac{1}{a}\int \frac{1}{y^2+k}\, dy.$$

  Then apply a trig substitution $y = \sqrt{k}\tan\theta$ to obtain

$$\frac{1}{a}\int \frac{1}{y^2+k}\, dy = \frac{\sqrt{k}}{a}\tan^{-1}\!\left(\frac{y}{\sqrt{k}}\right) = \frac{\sqrt{k}}{a}\tan^{-1}\!\left(\frac{x-h}{\sqrt{k}}\right).$$

Example Find $\int \frac{1}{(x+1)(x+2)^2} dx$.

Here, $P(x) = 1$ and $Q(x) = (x+1)(x+2)^2$. If I wanted to be sneaky, I could have asked for $\int \frac{1}{x^3+5x^2+8x+4} dx$ instead—which is the same question, but you'd need to do the factoring yourself.

According to the recipe outlined above, we must look for a split fraction of the form

$$\frac{1}{(x+1)(x+2)^2} = \frac{A}{x+1} + \frac{B}{x+2} + \frac{C}{(x+2)^2}.$$

To make the equation more explicit, let's add the fractions on the right. Set all of them to the least common denominator and add:

$$\begin{aligned}\frac{1}{(x+1)(x+2)^2} &= \frac{A}{x+1} + \frac{B}{x+2} + \frac{C}{(x+2)^2} \\ &= \frac{A(x+2)^2}{(x+1)(x+2)^2} + \frac{B(x+1)(x+2)}{(x+1)(x+2)^2} + \frac{C(x+1)}{(x+1)(x+2)^2} \\ &= \frac{A(x+2)^2 + B(x+1)(x+2) + C(x+1)}{(x+1)(x+2)^2}.\end{aligned}$$

The denominators are the same on both sides of the above equation, so we can focus our attention on the numerator:

$$A(x+2)^2 + B(x+1)(x+2) + C(x+1) = 1.$$

We can evaluate this equation for three different values of x to find the values of A, B, and C:

$$\begin{array}{cc} x=0 & 1 = 2^2A + 2B + C \\ x=-1 & 1 = A \\ x=-2 & 1 = -C \end{array}$$

so $A = 1$, $B = -1$, and $C = -1$. Thus,

$$\frac{1}{(x+1)(x+2)^2} = \frac{1}{x+1} - \frac{1}{x+2} - \frac{1}{(x+2)^2}.$$

We can now calculate the integral by integrating each of the terms:

$$\int \frac{1}{(x+1)(x+2)^2} dx = \ln(x+1) - \ln(x+2) + \frac{1}{x+2} + C.$$

Integration by parts

There is no general formula for finding the integral of the product of two functions $f(x)g(x)$. However, if one of the two functions in the product happens to look like the *derivative* of a function that we recognize, we can perform the following trick:

$$\int f(x)\, g'(x)\, dx \quad = \quad f(x)g(x) \quad - \int f'(x)g(x)\, dx.$$

This trick is called "integration by parts" and comes from the product rule for derivatives. We'll discuss how the formula is derived on page 377. First, let's see *why* we might want to use this trick.

The integration by parts procedure aims to simplify your task of integration. Both sides of the above equation involve an integral of the product of two functions: on the left we have $\int f(x)g'(x)dx$, while on the right we have $\int f'(x)g(x)dx$. The function $f(x)$ is replaced by its derivative $f'(x)$, while the function $g'(x)$ is replaced by its antiderivative function $g(x)$. Derivatives tend to simplify functions, whereas antiderivatives make functions more complicated. Thus, using integration by parts changes the integral calculation to one with a simplified $f(x)$.

It is easier to remember the integration by parts formula in its shorthand notation,

$$\int u\, dv = uv - \int v\, du.$$

You can think of integration by parts as a form of "double substitution," where you simultaneously replace u and dv. To be clear about what's happening during this substitution, I recommend you always make a little table like this:

$$\begin{aligned} u &= & \qquad\qquad dv &= \\ du &= & v &= . \end{aligned}$$

It's up to you to fill in the blanks. In the top row of this table, write the two factors from the original integral. Once you differentiate in the left column and integrate in the right column, the bottom row will contain the factors required for the integral on the right-hand side of the integration by parts formula.

Example 1 Find $\int xe^x \, dx$. We identify the good candidates for u and dv in the original expression, and follow the steps to apply the substitution:

$$\begin{aligned} u &= x & dv &= e^x \, dx, \\ du &= dx & v &= e^x. \end{aligned}$$

Next, apply the integration by parts formula,

$$\int u \, dv = uv - \int v \, du,$$

to obtain

$$\begin{aligned} \int xe^x \, dx &= xe^x - \int e^x \, dx \\ &= xe^x - e^x + C. \end{aligned}$$

Example 2 Find $\int x \sin x \ dx$. We choose the substitutions $u = x$ and $dv = \sin x dx$. With these choices, we have $du = dx$ and $v = -\cos x$. Integrating by parts gives us

$$\begin{aligned} \int x \sin x \, dx &= -x \cos x - \int (-\cos x) \ dx \\ &= -x \cos x + \int \cos x \ dx \\ &= -x \cos x + \sin x + C. \end{aligned}$$

Example 3 Often, you'll need to integrate by parts *multiple* times. To calculate $\int x^2 e^x \, dx$, we start by choosing the following substitutions:

$$\begin{aligned} u &= x^2 & dv &= e^x \ dx \\ du &= 2x \ dx & v &= e^x. \end{aligned}$$

After integration by parts, we have

$$\int x^2 e^x \, dx = x^2 e^x \; - \; 2 \int x e^x \, dx.$$

We apply integration by parts *again* to the remaining integral. This time, we use $u = x$ and $dv = e^x \, dx$, which gives $du = dx$ and $v = e^x$.

$$\begin{aligned} \int x^2 e^x \, dx &= x^2 e^x - 2 \int x e^x \, dx \\ &= x^2 e^x - 2 \left(x e^x - \int e^x \, dx \right) \\ &= x^2 e^x - 2 x e^x + 2 e^x + C. \end{aligned}$$

By now I hope you're starting to see why this integration by parts thing is good. If you remember to clearly write down the substitutions (indicating what is what in $\int udv$), and if you apply the formula correctly ($= uv - \int vdu$), you can break down any integral. Careful you don't make the wrong choice for your u and dv substitutions; if the integral $\int vdu$ is no simpler than the original $\int udv$, you are missing the point of integrating by parts, which is to make your life easier.

Sometimes, you might find yourself in a weird, self-referential loop when performing integration by parts. After a couple of integration by parts steps, it's possible to arrive at the very integral you started with! The way out of this loop is best shown by example.

Example 4 Evaluate the integral $\int \sin(x) e^x \;\; dx$. First, let $u = \sin(x)$ and $dv = e^x \; dx$, which gives $dv = \cos(x) dx$ and $v = e^x$. Using integration by parts,

$$\int \sin(x) e^x \, dx = e^x \sin(x) - \int \cos(x) e^x \, dx.$$

We integrate by parts again. This time, set $u = \cos(x)$, $dv = e^x dx$, and $du = -\sin(x)dx$, $v = e^x$. We obtain

$$\underbrace{\int \sin(x)e^x\,dx}_{I} \;=\; e^x\sin(x) - e^x\cos(x) \;\;-\;\; \underbrace{\int e^x\sin(x)\,dx}_{I}\,.$$

Do you see the Ouroboros? We could continue integrating by parts indefinitely in this way.

Let us clearly define what we are doing here. The question asks us to find I where

$$I = \int \sin(x)e^x\,dx,$$

and after completing two integration by parts steps, we obtain the equation

$$I = e^x\sin(x) - e^x\cos(x) - I.$$

Okay, good. Now move all the I's to one side:

$$2I = e^x\sin(x) - e^x\cos(x).$$

After factoring out e^x and dividing by 2 we finally obtain

$$\int \sin(x)e^x\,dx = I = \tfrac{1}{2}e^x\left(\sin(x) - \cos(x)\right) + C.$$

Integration by parts for integrals with limits

For definite integrals, the integration by parts rule must account for evaluation at the function's limits:

$$\int_a^b u\,dv = (uv)\Big|_a^b \;-\; \int_a^b v\,du.$$

This expression tells us to evaluate the change in the product uv at the limits of integration.

Example 5 Find $\int_0^5 xe^x\,dx$. We've already seen this example, but this time it includes limits of integration. The first part of the procedure is the same. We apply the substitution

$$\begin{aligned} u &= x & dv &= e^x\,dx, \\ du &= dx & v &= e^x. \end{aligned}$$

Then use the formula for integration by parts with limits:

$$\begin{aligned} \int_0^5 xe^x\,dx &= (xe^x)\Big|_0^5 \;-\; \int_0^5 e^x\,dx \\ &= (xe^x)\Big|_0^5 \;-\; e^x\Big|_0^5 \\ &= \left[5e^5 - 0e^0\right] \;-\; \left[e^5 - e^0\right] \\ &= 5e^5 - e^5 + 1 \\ &= 4e^5 + 1. \end{aligned}$$

Derivation of the Integration by parts formula

Remember the product rule for derivatives?

$$\frac{d}{dx}(f(x)g(x)) = \frac{df}{dx}g(x) + f(x)\frac{dg}{dx}.$$

We can rewrite it as

$$f(x)\frac{dg}{dx} = \frac{d}{dx}(f(x)g(x)) \;-\; \frac{df}{dx}g(x).$$

Take the integral on both sides of the equation:

$$\begin{aligned} \int \left[f(x)\frac{dg}{dx}\right] dx \;&=\; \int \left[\frac{d}{dx}(f(x)g(x)) - \frac{df}{dx}g(x)\right] dx \\ \int f(x)\frac{dg}{dx}\;dx \;&=\; \int \left[\frac{d}{dx}(f(x)g(x))\right] dx \;-\; \int \frac{df}{dx}g(x)dx\,. \end{aligned}$$

At this point, think back to the fundamental theorem of calculus (see page 346), which says the derivative and the integral are inverse operations, $\int \left(\frac{d}{dx}h(x)\right) dx = h(x)$.

We apply this logic to simplify the first term on the right-hand side of the above equation and obtain

$$\int f(x)\frac{dg}{dx}\,dx \;=\; f(x)g(x) \;\;-\;\; \int \frac{df}{dx}g(x)\,dx,$$

which is the integration by parts formula.

Outro

We are done. You know all the integration techniques. I know it took a while, but we had to discuss a lot of tricks. In any case, my job of teaching you is done. Now *your* job begins. Practice all the examples you can find. Try solving all the exercises in the back of the book. You must practice the tricks until you become comfortable with them.

Here's a suggestion for you. Make your own trophy case for formulas. As you cover ground in your homework assignments, create and maintain a formula-sheet where you record any complex integrals you have personally calculated from first principles. By the end of the class, if your trophy case contains 50 integrals you have personally calculated all by yourself, then you will earn 100% on your final. Another thing to try: review the integral formulas in the back of the book and see how many of them you can derive.

Links

[More examples of integration techniques]
http://en.wikibooks.org/wiki/Calculus/Integration_techniques/

5.16 Applications of integration

Applications to mechanics

Calculus was essentially invented *for* mechanics, so it's not surprising there are many links between the two subjects.

Kinematics

Suppose a constant force F_{net} is applied to an object of mass m. Newton's second law tells us the acceleration of the object will be constant and equal to $a = \frac{F_{\text{net}}}{m}$.

We can find the equation of motion for the object $x(t)$ by integrating $a(t)$ twice, since $a(t) = x''(t)$. We start with the acceleration function $a(t) = a$ and integrate once to obtain

$$v(\tau) = \int_0^\tau a(t)\, dt = at + v_i,$$

where $v_i = v(0)$ is the object's initial velocity at $t = 0$. We obtain the position function by integrating the velocity function and adding the initial position $x_i = x(0)$:

$$x(\tau) = \int v(t)\, dt = \int (at + v_i)\, dt = \frac{1}{2}a\tau^2 + v_i\tau + x_i.$$

Non-constant acceleration

If the net force acting on the object is not constant, the acceleration will not be constant either. In general, both force and mass can change over time, and if they do, acceleration will also change over time $a(t) = \frac{F_{\text{net}}(t)}{m(t)}$. This sort of problem is usually not covered in a first mechanics course because the establishment assumes it is too complicated for you to handle.

Now that you know more about integrals, you can learn how to predict the motion of an object for an arbitrary acceleration function $a(t)$. To find the velocity at time $t = \tau$, we must sum all acceleration felt by the object between $t = 0$ and $t = \tau$:

$$v(\tau) = v_i + \int_0^\tau a(t)\, dt.$$

The equation of motion $x(t)$ is obtained by integrating the velocity $v(t)$:

$$x(s) = x_i + \int_0^s v(\tau)\, d\tau = \int_0^s \left[v_i + \int_0^\tau a(t)\, dt \right] d\tau.$$

The above expression looks quite intense, but in fact it is no more complicated than the simple integrals used in UAM. The expression *looks* complicated because it contains three different variables representing time, as well as two consecutive integration steps.

Gravitational potential energy

The work done by a force $\vec{F}$ during a displacement $\vec{d}$ is computed using the integral $W = \int_0^d \vec{F} \cdot d\vec{x}$, which simplifies to $W = \vec{F} \cdot \vec{d}$ for a constant force. The negative of the work done by a conservative force defines the *potential energy* function associated with that force.

Since gravity $\vec{F}_g$ is a conservative force, we can integrate it to obtain the gravitational potential energy U_g. On the surface of the Earth, $\vec{F}_g = -gm\hat{\jmath}$. The negative sign means the force of gravity acts in the direction opposite to "upward," as represented by the $\hat{\jmath}$ unit vector pointing in the positive y-direction (toward the sky). In particular, gravitational force as a function of height $\vec{F}_g(y)$ is a constant $\vec{F}_g(y) = \vec{F}_g$. By definition, gravitational potential energy is the negative of the integral of the force over some distance, say from height $y_i = 0$ to height $y_f = h$:

$$\Delta U_g = U_g(h) - U_g(0) = -\int_{y_i}^{y_f} \vec{F}_g \cdot d\vec{y} = -\int_0^h -mg\, dy = \Big[mgy \Big]_0^h = mgh.$$

The general form of the gravitational force acting on an object of mass m due to another object of mass M is given by Newton's famous one-over-r-squared law,

$$\vec{F}_g = \frac{GMm}{r^2}\hat{r}.$$

In this law, r is the distance between the objects and $\hat{r}$ points toward the other object.

The general formula for gravitational potential energy is obtained by taking the integral of the gravitational force over some distance. Imagine a planet of mass m and another planet of mass M. The two masses start infinitely far away from each other and slowly move closer until they are a distance $r = R$ apart. The change in gravitational potential from $r = \infty$ to $r = R$ is

$$\begin{aligned}\Delta U_g(R) &= \int_{r=\infty}^{r=R} \frac{GMm}{r^2}\, dr \\ &= GMm \int_{\infty}^{R} \frac{1}{r^2}\, dr \\ &= GMm \left[\frac{-1}{r}\right]_{\infty}^{R} \\ &= GMm \left[\frac{-1}{R} - \frac{-1}{\infty}\right] \\ &= -\frac{GMm}{R}\,.\end{aligned}$$

The gravitational potential energy of two planets a distance R apart is negative since work is *required* to pull the planets apart once they have come together.

There is an important physics lesson to learn here. For each conservative force $\vec{F}_?(x)$, there is an associated potential energy function $U_?(x)$ that is defined as the negative of the work done when moving an object against the force $\vec{F}_?(x)$:

$$\text{Given } \vec{F}_?(x) \qquad \Rightarrow \qquad U_?(x) \equiv -\int_0^x \vec{F}_?(u)\cdot d\vec{u}.$$

We can use this relationship in the other direction too. Given a potential energy function $U_?(x)$, we can find the force $F_?(x)$ associated with that potential energy function by taking the derivative:

$$\text{Given } U_?(x) \qquad \Rightarrow \qquad F_?(x) \equiv -\frac{d}{dx}\Big[U_?(x)\Big].$$

The negative of the derivative of the gravitational potential energy $U_g(y) = mgy$ gives $F_g(y) = -mg$. The negative of the derivative of the spring potential energy $U_s(x) = \frac{1}{2}kx^2$ gives $F_s(x) = -kx$.

Integrals over circular objects

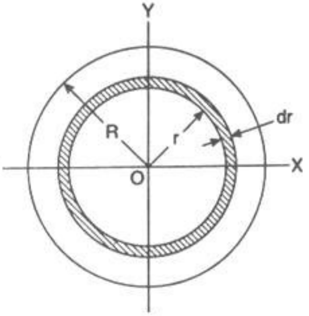
We can use integration to calculate the area and volume formulas of objects with circular symmetries.

Consider the disk-shaped region described by the equation $D = \{x, y \in \mathbb{R} \mid x^2 + y^2 \leq R^2\}$. In polar coordinates we describe this region as $r \leq R$, where it is implicit that the angle θ varies between 0 and 2π. Because this region is two-dimensional, computing an integral over this region requires a *double integral*, which is the subject of multivariable calculus. Even before you learn about double integrals, you still know enough to integrate over a circular region by breaking the region into thin circle-shaped slices of disk dD.

Similar to the way a horizontal slice through an onion consists of many thin onion rings, we can break the disk into a number of thin circular strips of width dr. The circular strip with radius r has an area of

$$dD = 2\pi r\, dr,$$

since $2\pi r$ is the circumference of a circle with radius r, and since the width of the strip is dr.



Using this method to break apart the disk, we can check that adding the areas of all the "pieces of disk"

dD gives a total area of πR^2:

$$A_{\text{disk}} = \int_D dD = \int_{r=0}^{r=R} 2\pi r\ dr = 2\pi \int_0^R r\ dr = 2\pi \left[\frac{r^2}{2}\right]_0^R = \pi R^2.$$

The following sections discuss different variations of this breaking-an-onion-slice-into-onion-rings idea. We can use the circular symmetry of various objects to compute their area, volume, and moment of inertia.

Total mass of a disk

Suppose you have a disk of total mass m and radius R. You can think of the disk as being made of parts, each of mass dm, such that adding all the parts gives the disk's total mass:

$$\int_{\text{disk}} dm = m.$$

The mass density is defined as the total mass divided by the area of the disk: $\sigma = \frac{m}{A_{\text{disk}}} = \frac{m}{\pi R^2}$[kg/m^2]. Mass density corresponds to the amount of mass per unit area. We can split the disk into concentric circular strips of width dr. The mass contribution of each strip is equal to σ times the area of that strip. The strip at radius r has circumference $2\pi r$ and width dr, so its mass contribution is $dm = \sigma 2\pi r\ dr$.

Let's check that we obtain the total mass by adding the pieces:

$$\int_{\text{disk}} dm = \int_0^R \sigma 2\pi r\ dr = 2\pi\sigma \left[\frac{r^2}{2}\right]_0^R = 2\pi \frac{m}{\pi R^2}\frac{R^2 - 0}{2} = m.$$

Moment of inertia of a disk

An object's moment of inertia is a measure of how difficult it is to make the object turn. The moment of inertia appears in the rotational version of $F = ma$, in place of the mass m: $\mathcal{T} = I\alpha$. The moment of inertia I also appears in the formula for angular momentum $L = I\omega$ and the formula for rotational kinetic energy $K_r = \frac{1}{2}I\omega^2$.

To compute an object's moment of inertia, we must add all the mass contributions dm, multiplying each by r^2, where r is the distance of the piece dm from the disk's centre:

$$I = \int_{\text{disk}} r^2 \, dm.$$

We can perform the integral over the whole disk by adding the contributions of all the strips:

$$I_{\text{disk}} = \int_0^R r^2 \, dm = \int_0^R r^2 \sigma 2\pi r \, dr = \int_0^R r^2 \frac{m}{\pi R^2} 2\pi r \, dr =$$
$$= \frac{2m}{R^2} \int_0^R r^3 \, dr = \frac{2m}{R^2} \left[\frac{r^4}{4} \right]_0^R = \frac{2m}{R^2} \frac{R^4}{4} = \frac{1}{2} m R^2.$$

Arc lengths of a curve

Given a function $y = f(x)$, how can you calculate the total *length* ℓ of the graph of $f(x)$ between $x = a$ and $x = b$?

If $f(x)$ is the equation of a line, the length of its graph can be calculated as the hypotenuse of the change-in-x and the change-in-y triangle: $\ell = \sqrt{\text{run}^2 + \text{rise}^2} = \sqrt{(b-a)^2 + (f(b) - f(a))^2}$.

However, if the function is *not* a straight line, we need to apply this hypotenuse calculation to each piece of the curve $d\ell = \sqrt{dx^2 + dy^2}$, then add all the contributions as an integral.

The arc length ℓ of a graph $y = f(x)$ on the interval $x \in [a, b]$ is

$$\ell = \int d\ell = \int \sqrt{dx^2 + dy^2} = \int \sqrt{\left(1 + \frac{dy^2}{dx^2}\right) dx^2} = \int_a^b \sqrt{1 + (f'(x))^2} dx.$$

Surface of revolution

We modify the arc-length formula to calculate the surface area A of a solid of revolution. An object with circular symmetry can be generated by a revolution of some curve $f(x)$ around the x-axis.

Each piece of length $d\ell$ must be multiplied by $2\pi f(x)$, since each piece rotates around the x-axis in a circle of radius $f(x)$. The area of the surface of revolution traced by $f(x)$ as it rotates around the x-axis is given by the following integral:

$$A = \int 2\pi f(x) d\ell = \int_{x_i}^{x_f} 2\pi f(x)\, \sqrt{1 + (f'(x))^2}\; dx.$$

Volumes of revolution

Let's move on to three-dimensional integrals, or integrals over volumes. Again, we'll use the circular symmetry of the object's volume to split the object into little "pieces of volume" and then compute an integral to find the total volume.

Disk method

We can describe the volume of an object with circular symmetry as the sum of a number of disks. Each disk will have thickness dx and a radius proportional to the function $f(x)$. In other words, the function $f(x)$ describes the object's outer boundary. The area of each disk is $\pi(f(x))^2$ and its thickness is dx.

The volume of a solid of revolution with boundary $f(x)$ rotated around the x-axis is given by the formula

$$V = \int A_{\text{disk}}\, dx = \int \pi (f(x))^2\, dx.$$

Example Use the disk method to calculate the volume of a sphere with radius R. The volume we want to calculate is bounded by the curve $f(x) = \sqrt{R^2 - x^2}$ and the horizontal line $y = 0$. Our limits of integration are the x-values where the curve intersects the line $y = 0$, namely, $x = \pm R$. We have

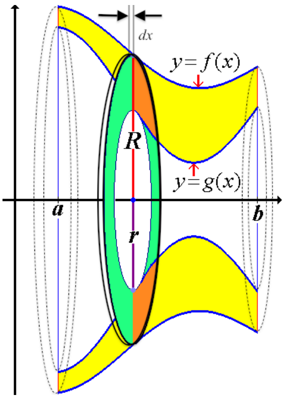
$$\begin{aligned} V_{\text{sphere}} &= \int_{-R}^{R} \pi (R^2 - x^2) dx \\ &= \pi \Big(\int_{-R}^{R} R^2\, dx - \int_{-R}^{R} x^2\, dx \Big) \end{aligned}$$

$$
\begin{aligned}
&= \pi\left(R^2 x\Big|_{-R}^{R} - \frac{x^3}{3}\Big|_{-R}^{R} \right) \\
&= \pi\left(R^2\left[R-(-R)\right] - \left[\frac{R^3}{3} - \frac{(-R)^3}{3}\right] \right) \\
&= \pi\left(2R^3 - \frac{2R^3}{3} \right) \\
&= \pi\frac{6R^3 - 2R^3}{3} \\
&= \frac{4\pi R^3}{3}\,.
\end{aligned}
$$

Indeed, this is the formula for the volume of a sphere that we first encountered in Chapter 1 (see page 92).

Washer method

The washer method is a generalization of the disk method for computing volumes. Consider a volume of revolution with an inner radius described by the function $g(x)$ and an outer radius described by the function $f(x)$. The diagram on the right shows such a volume of revolution. Instead of using thin disks, we can represent this volume as the sum of thin washers, which are disks of radius $f(x)$ with a middle section of radius $g(x)$ removed.



The volume dV of each washer is equal to the area $\pi(f(x))^2$ minus the removed area $\pi(g(x))^2$, times the thickness dx. The total volume is given by

$$
\begin{aligned}
V &= \int dV \\
&= \int A_{\text{washer}}\,dx \quad = \int \pi[(f(x))^2 - (g(x))^2]\,dx.
\end{aligned}
$$

Cylindrical shell method

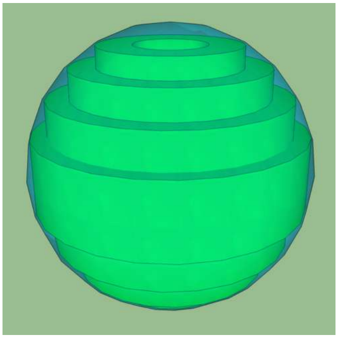
We can split any circularly symmetric volume into thin, cylindrical shells of thickness dr. If the volume has a circular symmetry and is bounded from above by $F(r)$ and from below by $G(r)$, then the integral over the volume will be

$$\begin{aligned} V &= \int C_{\text{shell}}(r)\, h_{\text{shell}}(r)\, dr \\ &= \int_a^b 2\pi r|F(r) - G(r)|\, dr. \end{aligned}$$

The cylindrical shell with radius r has circumference $2\pi r$, thickness dr, and its height is described by the expression $|F(r) - G(r)|$.

Example Calculate the volume of a sphere of radius R using the cylindrical shell method. We are talking about the region enclosed by the surface $x^2+y^2+z^2 = R^2$.

At radius $r = \sqrt{x^2+y^2}$, the cylindrical shell will be bounded from above by $z = F(r) = \sqrt{R^2-r^2}$, and bounded from below by $z = G(r) = -\sqrt{R^2-r^2}$. The circumference of the shell is $2\pi r$ and its width is dr. The integral proceeds like this:



$$\begin{aligned} V &= \int_0^R 2\pi r|F(r) - G(r)|\, dr \\ &= \int_0^R 2\pi r 2\sqrt{R^2-r^2}\, dr \\ &= -2\pi \int_{R^2}^0 \sqrt{u}\, du \qquad (\text{using } u = R^2 - r^2, du = -2rdr) \\ &= -2\pi \left[\tfrac{2}{3}u^{3/2}\right]_{R^2}^0 = -2\pi[0 - \tfrac{2}{3}R^3] = \frac{4\pi R^3}{3}. \end{aligned}$$

Exercises

Exercise 1 Calculate the volume of a cone with radius R and height h that is generated by the revolution of the region bounded by $y = R - \frac{R}{h}x$ and the lines $y = 0$ and $x = 0$ around the x-axis. Ans: $\frac{\pi R^2 h}{3}$

Exercise 2 Calculate the volume of the solid of revolution generated by revolving the region bounded by the curve $y = x^2$ and the lines $x = 0$, $x = 1$, and $y = 0$ around the x-axis. Ans:$\frac{\pi}{5}$

Exercise 3 Calculate the volume of the solid of revolution generated by revolving the region bounded by the curves $y = x^2$ and $y = x^3$ and the lines $x = 0$ and $x = 1$ around the x-axis. Ans: $\frac{2\pi}{35}$

Exercise 4 Find the volume of a vertical cone with radius R and height h formed by the revolution of the region bounded by the curves $y = h - \frac{h}{R}x$, $y = 0$ and $x = 0$, around the y-axis. Use the cylindrical shell method. Ans:$\frac{\pi R^2 h}{3}$

Links

[An animation showing how a volume of revolution is constructed]
`http://mathforum.org/mathimages/index.php/Volume_of_Revolution`

5.17 Improper integrals

Imagine you want to find the area under the function $f(x) = \frac{1}{x^2}$ from $x = 1$ all the way to $x = \infty$. This kind of calculation is known as an *improper integral* since one of the endpoints of the integration is not a regular number, but infinity.

We can compute this integral as the following limit:

$$\int_1^\infty \frac{1}{x^2}\,dx \equiv \lim_{b\to\infty}\int_1^b \frac{1}{x^2}\,dx = \lim_{b\to\infty}\left[\frac{-1}{x}\right]_1^b = \lim_{b\to\infty}\left[-\frac{1}{b}+\frac{1}{1}\right] = 1.$$

This calculation describes an integration over a region with infinite width. Because the height of the region ($f(x) = \frac{1}{x^2}$) becomes smaller and smaller, the region still has finite area.

Definitions

An improper integral is an integral in which one of the limits of integration goes to infinity. Improper integrals are evaluated as regular integrals, where infinity is replaced by a dummy variable, after which a limit calculation is applied to take the dummy variable to infinity:

$$\int_a^\infty f(x)\,dx \equiv \lim_{b\to\infty}\int_a^b f(x)\,dx = \lim_{b\to\infty}[F(b) - F(a)],$$

where $F(x)$ is the antiderivative function of $f(x)$.

Applications

Later in this chapter, we'll learn about the "integral test" for the convergence of series, which requires the evaluation of an improper integral.

5.18 Sequences

A sequence is an ordered list of numbers, and usually follows some pattern, much like "find the pattern" questions on IQ tests. We can study the properties of sequences as mathematical objects. For example, by checking whether the sequence *converges* to some limit.

Understanding sequences is a prerequisite for understanding series, which is an important topic we will discuss in the next section.

Definitions

- $\mathbb{N}$: the set of *natural* numbers $\mathbb{N} \equiv \{0, 1, 2, 3, \ldots\}$
- $\mathbb{N}_+ = \mathbb{N} \backslash \{0\}$: the set of *strictly positive* natural numbers $\{1, 2, 3, \ldots\}$. The set $\mathbb{N}_+$ is the same as $\mathbb{N}$, except $\mathbb{N}_+$ starts from 1 instead of 0.
- a_n: a sequence of numbers $(a_0, a_1, a_2, a_3, a_4, \ldots)$. You can also think of each sequence as a function

  $$a : \mathbb{N} \to \mathbb{R},$$

  where the input n is an integer (the *index* into the sequence) and the output is some number $a_n \in \mathbb{R}$.

Examples of sequences

Consider the following common sequences.

Arithmetic progression

A sequence is an arithmetic progression if the terms of the sequence differ by a constant amount. The terms in the simplest arithmetic progression differ by one:

$$(0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ \ldots).$$

This sequence is described by the formula

$$a_n = n, \qquad n \in \mathbb{N}.$$

More generally, an arithmetic sequence can start at any value a_0 and make jumps of size d at each step:

$$a_n = a_0 + nd, \qquad n \in \mathbb{N}.$$

Harmonic sequence

In a *harmonic* sequence, each element of the sequence is inversely proportional to its index n:

$$\left(1,\ \frac{1}{2},\ \frac{1}{3},\ \frac{1}{4},\ \frac{1}{5},\ \frac{1}{6},\ \ldots\right)$$

$$a_n = \frac{1}{n}, \qquad n \in \mathbb{N}_+.$$

More generally, we can define a p-sequence in which the index n appears in the denominator raised to the power p:

$$a_n = \frac{1}{n^p}, \qquad n \in \mathbb{N}_+.$$

For example, when $p = 2$, the result is a sequence of inverse squares of the integers:

$$\left(1,\ \frac{1}{4},\ \frac{1}{9},\ \frac{1}{16},\ \frac{1}{25},\ \frac{1}{36},\ \ldots\right).$$

Geometric sequence

By using the index as the exponent of a fixed number r, we obtain the geometric series

$$a_n = r^n, \quad n \in \mathbb{N},$$

which is a sequence of the form

$$\left(1, r, r^2, r^3, r^4, r^5, r^6, \ldots\right).$$

If we choose $r = \frac{1}{2}$, then the geometric series with this ratio will be

$$\left(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \ldots\right).$$

Fibonacci

The Fibonacci numbers are constructed according to the following pattern. The first Fibonacci number is 0, the second Fibonacci number is 1, and each subsequent number is the sum of the two preceding it:

$$a_0 = 0, a_1 = 1, \qquad a_n = a_{n-1} + a_{n-2}, \quad n > 2.$$

$$(0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \ldots).$$

Convergence

We say a sequence a_n *converges* to a limit L, written mathematically as

$$\lim_{n\to\infty} a_n \; = \; L,$$

if for large values of n the terms in the sequence become arbitrarily close to the value L.

More precisely, the limit expression $\lim\limits_{n\to\infty} a_n \; = \; L$ means that for *any* precisions $\epsilon > 0$, we can pick a number N_ϵ such that

$$|a_n - L| < \epsilon, \qquad \forall n \geq N_\epsilon.$$

The notion of a limit of a sequence is the same as the notion of a limit of a function. Just as we learned how to calculate which number the function $f(x)$ tends to for large x values (see page 285), we can study which number the sequence a_n tends to for large n values.

Ratio convergence

The numbers in the Fibonacci sequence grow indefinitely large ($\lim\limits_{n\to\infty} a_n \; = \; \infty$), while the ratio of $\frac{a_{n+1}}{a_n}$ converges to a constant:

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \varphi = \frac{1+\sqrt{5}}{2} \approx 1.618033\ldots$$

This constant is known as the *golden ratio*.

Calculus on sequences

If a sequence a_n is like a function $f(x)$, we should be able to perform calculus on it. We already saw how we can take limits of sequences, but can we also compute derivatives and integrals of sequences? Derivatives are a no-go, because they depend on the function $f(x)$ being *continuous*, and sequences are only defined for integer values. We *can* take integrals of sequences, however, and this is the subject of the next section.

5.19 Series

Can you compute $\ln(2)$ using only a basic calculator with four operations, $\boxed{+}$, $\boxed{-}$, $\boxed{\times}$, and $\boxed{\div}$? I can tell you one way. Compute the following infinite sum:

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots .$$

Since the sum is infinite, it will take a while to obtain the value of $\ln(2)$, but if you keep adding more terms in the sum, you will eventually obtain the answer $\ln(2) = 0.693147\ldots$.

Let's make the computer carry out the summation for us. First we define the formula for the n^{th} term in the series $a_n = \frac{(-1)^{n+1}}{n}$, then we compute the sum of the first 100, 1000, and 1000000 terms:

```
>>> def an_ln2(n): return 1.0*(-1)**(n+1)/n
>>> sum([ an_ln2(n)  for n in range(1,100) ])
      0.69
>>> sum([ an_ln2(n)  for n in range(1,1000) ])
      0.693
>>> sum([ an_ln2(n)  for n in range(1,1000000) ])
      0.693147
```

Observe how the approximation becomes more accurate as more terms are added in the sum. A lot of practical mathematical computations are performed in this *iterative* fashion. In this section we'll learn about a powerful technique for calculating quantities to arbitrary precision by summing together more and more terms of a series.

Definitions

- $\mathbb{N} \equiv \{0, 1, 2, 3, 4, 5, 6, \ldots\}$: the set of natural numbers
- $\mathbb{N}_+ \equiv \mathbb{N} \setminus \{0\} \equiv \{1, 2, 3, 4, 5, 6, \ldots\}$: the set of positive natural numbers
- a_n: a sequence of numbers $(a_0, a_1, a_2, a_3, a_4, \ldots)$
- $\sum$: sum. This symbol indicates taking the sum of several objects grouped together. The summation sign is the short way to express certain long expressions:

$$a_3 + a_4 + a_5 + a_6 + a_7 = \sum_{3 \le n \le 7} a_n = \sum_{n=3}^{7} a_n.$$

- $\sum a_n$: the series a_n is the sum of all terms in the sequence a_n:

$$S_\infty = \sum_{i=1}^{\infty} = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + \cdots .$$

- $n!$: the *factorial* function $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$, if $n \ge 1$. We define $0! = 1$.
- $f(x) = \sum_{n=0}^{\infty} c_n x^n$: the *Taylor series* approximation of the function $f(x)$. It has the form of an infinitely long polynomial $c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots$ where the coefficients c_n are chosen so as to encode the properties of the function $f(x)$.

Exact sums

Formulas exist for calculating the exact sum of certain series.

The sum of the geometric series of length N is

$$S_N = \sum_{n=0}^{N} r^n = 1 + r + r^2 + \cdots + r^N = \frac{1 - r^{N+1}}{1 - r}.$$

If $|r| < 1$, taking the limit $N \to \infty$ in the above expression leads to

$$S_\infty = \lim_{N\to\infty} S_N = \sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + \cdots = \frac{1}{1 - r}.$$

Example Consider the geometric series with $r = \frac{1}{2}$. Applying the above formula, we obtain

$$S_\infty = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots = \frac{1}{1 - \frac{1}{2}} = 2.$$

You can also visualize this infinite summation graphically. Imagine starting with a piece of paper of size one-by-one, then adding next to it a second piece of paper with half the size of the first, and a third piece with half the size of the second, and so on. The total area occupied by these pieces of papers is shown in Figure 5.13.

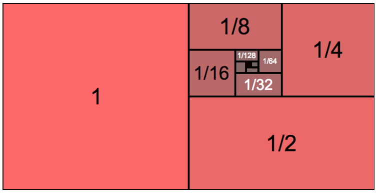


Figure 5.13: A graphical representation of the infinite sum of the geometric series with $r = \frac{1}{2}$. The area of each region corresponds to one of the terms in the series. The total area is equal to $\sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n = \frac{1}{1-\frac{1}{2}} = 2$.

We'll now state without proof a number of other formulas where the sum of a series can be obtained as a closed-form expression.

The formulas for the sum of the first N positive integers and the sum of the squares of the first N positive integers are

$$\sum_{n=1}^{N} n = \frac{N(N+1)}{2}, \qquad \sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}.$$

The sum of the first $N+1$ terms in an arithmetic sequence is

$$\sum_{n=0}^{N}(a_0 + nd) = a_0(N+1) + \frac{N(N+1)}{2}d.$$

Another group of series with exact formulas for their sums are the p-series involving even values of p:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}, \quad \sum_{n=1}^{\infty} \frac{1}{n^6} = \frac{\pi^6}{945}.$$

Other closed-form sums include:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \ln(2), \quad \sum_{n=1}^{\infty} \frac{1}{4n^2-1} = \frac{1}{2},$$

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^3} = \frac{\pi^3}{32}, \quad \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{\pi^4}{96}.$$

Convergence and divergence of series

Even when we cannot compute an exact expression for the sum of a series, it is important to distinguish series that converge from series that do not converge.

We say a series $\sum a_n$ *converges* if the infinite sum $S_\infty \equiv \sum_{n\in\mathbb{N}} a_n$ equals some finite number $L \in \mathbb{R}$.

$$S_\infty = \sum_{n=0}^{\infty} a_n = L \quad \Rightarrow \quad \text{the series } \sum a_n \text{ converges.}$$

If the infinite sum $S_\infty \equiv \sum_{n\in\mathbb{N}} a_n$ grows to infinity, we say the series $\sum a_n$ *diverges.*

$$S_\infty = \sum_{n=0}^{\infty} a_n = \pm\infty \quad \Rightarrow \quad \text{the series } \sum a_n \text{ diverges.}$$

Convergence of a series is not the same as convergence of the underlying sequence a_n. Consider the sequence of partial sums $S_N = \sum\limits_{n=0}^{N} a_n$:

$$S_0, S_1, S_2, S_3, \ldots,$$

where each of the terms in the sequence corresponds to

$$a_0, \quad a_0 + a_1, \quad a_0 + a_1 + a_2, \quad a_0 + a_1 + a_2 + a_3, \ldots.$$

We say the series $\sum a_n$ converges if the sequence of partial sums S_N converges to a limit L:

$$\lim_{N\to\infty} S_N = L.$$

This limit statement indicates that the partial sums S_N approach the number L as we include more terms in the series.

The precise meaning of the limit statement is as follows. For any precision $\epsilon > 0$, there exists a starting point N_ϵ such that, for all $N > N_\epsilon$, it will be true that

$$|S_N - L| < \epsilon.$$

The number N_ϵ corresponds to how many terms of the series you need for the partial sum S_N to become ϵ-close to the limit of the series L.

Convergence tests

A great deal of what you need to know about series involves learning different *tests* you can perform to check whether a series converges or diverges. We'll now discuss a number of these tests.

Sequence convergence test

The only way the infinite sum $\sum_{n=0}^{\infty} a_n$ will converge is if the elements of the sequence a_n tend to zero for large n. This observation gives us a simple series *divergence* test. If $\lim_{n\to\infty} a_n \neq 0$ then $\sum_{n=0}^{\infty} a_n$ diverges. How could an infinite sum of non-zero quantities add to a finite number?

Absolute convergence

If $\sum_n |a_n|$ converges, $\sum_n a_n$ also converges. The opposite is not necessarily true, since the convergence of a_n might be due to negative terms *cancelling* with positive terms.

A sequence a_n for which $\sum_n |a_n|$ converges is called *absolutely convergent*. A sequence b_n for which $\sum_n b_n$ converges but $\sum_n |b_n|$ diverges is called *conditionally convergent*.

Decreasing alternating sequences

An alternating series a_n in which the absolute values of the terms is decreasing ($|a_n| > |a_{n+1}|$), and tend to zero ($\lim a_n = 0$) converges. For example, we know the series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots$ converges because it is a decreasing alternating series and $\lim_{n\to\infty} \frac{1}{n} = 0$.

Integral test

If the integral $\int_a^{\infty} f(x)\,dx$ is finite, then the series $\sum_n f(n)$ converges. If the integral $\int_a^{\infty} f(x)\,dx$ diverges, then the series $\sum_n f(n)$ also diverges.

The improper integral is defined as the limit expression:

$$\int_a^{\infty} f(x)\,dx \equiv \lim_{b\to\infty} \int_a^b f(x)\,dx.$$

The p-series converges if $\mathbf{p > 1}$

The convergence conditions for p-series, $a_n = \frac{1}{n^p}$, can be obtained using the integral test.

The series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if $p > 1$, and diverges if $p \leq 1$. Note that $p = 1$ corresponds to the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ which diverges.

Limit comparison test

Suppose $\lim\limits_{n\to\infty} \dfrac{a_n}{b_n} = L$. We can draw the following conclusions:

- If $0 < L < \infty$, then $\sum_n a_n$ and $\sum_n b_n$ either both converge or both diverge.
- If $L = 0$, and if $\sum_n b_n$ converges, then $\sum_n a_n$ also converges.
- If $L = \infty$, and if $\sum_n b_n$ diverges, then $\sum_n a_n$ also diverges.

The n^th root test

If r is defined by $r = \lim\limits_{n\to\infty} \sqrt[n]{|a_n|}$, then $\sum_n a_n$ diverges if $r > 1$ and converges if $r < 1$. If $r = 1$, the test is inconclusive.

The ratio test

The most useful convergence test is the ratio test. To use the ratio test, compute the limit of the ratio of successive terms in the sequence:

$$R = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|.$$

The series $\sum_n^{\infty} a_n$ converges if $R < 1$, and $\sum_n^{\infty} a_n$ diverges if $R > 1$. If $R = 1$, the test is inconclusive.

Taylor series

The *Taylor series* of a function $f(x)$ approximates the function by an infinitely long polynomial:

$$f(x) = \sum_{i=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \cdots .$$

Each term in the series is of the form $a_n = c_n x^n$, where the coefficient c_n depends on the properties of the function $f(x)$. For example, the Taylor series of the function $\sin(x)$ is

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots .$$

How do the coefficients c_n depend on the function $f(x)$? How can we compute the Taylor series for other functions?

The general procedure for computing the coefficients c_n in the Taylor series of a function $f(x)$ is to choose c_n equal to the n^{th} derivative of $f(x)$ divided by $n!$:

$$\begin{aligned} f(x) &= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n. \end{aligned}$$

Using this formula and your knowledge of derivatives, you can compute the Taylor series of any function $f(x)$.

Example Find the Taylor series of $f(x) = e^x$. The formula for the n^{th} coefficient in the Taylor series of the function $f(x)$ is $c_n = \frac{f^{(n)}(0)}{n!}$. The first derivative of $f(x) = e^x$ is $f'(x) = e^x$. The second derivative of $f(x) = e^x$ is $f''(x) = e^x$. In fact, all the derivatives of $f(x)$ will be e^x because the e^x is a special function that is equal to its derivative! The n^{th} coefficient in the power series of $f(x) = e^x$ at the point $x = 0$ is equal to the value of the n^{th} derivative of $f(x)$ evaluated at $x = 0$. In the case of $f(x) = e^x$ we have $f^{(n)}(0) = e^0 = 1$, so the coefficient of the n^{th} term is $c_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{n!}$. The Taylor series of $f(x) = e^x$ is

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

Here are the Taylor series of some other commonly used functions:

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\, x^{2n}$$

$$\ln(x+1) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}\, x^n$$

$$\cosh(x) = 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \frac{x^6}{6!} + \frac{x^8}{8!} + \frac{x^{10}}{10!} + \cdots = \sum_{n=0}^{\infty} \frac{1}{(2n)!}\, x^{2n}$$

$$\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \frac{x^9}{9!} + \frac{x^{11}}{11!} + \cdots = \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} x^{2n+1}$$

Note the similarities between the power series of $\cos$ and $\cosh$. The formulas are the same, but the hyperbolic version does not alternate. When the formula for the n^{th} coefficient c_n contains the factor $(-1)^n$, the terms in the series will alternate between positive and negative.

Both $\cos$ and $\cosh$ are *even* functions, meaning $f(x) = f(-x)$. The "evenness" of $\cos$ and $\cosh$ can also be confirmed by comparing their power series, which contain only even powers of x. Note how the index $n \in \mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$ is transformed to an even index $2n \in \{0, 2, 4, 6, 8, \ldots\}$. Similarly, we use the index $(2n + 1)$ to obtain only odd numbers $(2n + 1) \in \{1, 3, 5, 7, \ldots\}$.

Terminology

A more specific mathematical term for the series we discussed above is *Maclaurin series*, which is a specific case of a Taylor series.

The coefficients c_n in the *Taylor series* of $f(x)$ are obtained by computing the value of the n^{th} derivative of $f(x)$. The Taylor series of a function $f(x)$ at the point $x = a$ is given by

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n. \end{aligned}$$

The *Maclaurin series* of $f(x)$ is the Taylor series of $f(x)$ with $a = 0$:

$$\begin{aligned} f(x) &= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n. \end{aligned}$$

The term *power series* is also used to describe Taylor series, since each term in the series contains x raised to a certain power.

Radius of convergence for power series

Consider the power series $f(x) = \sum c_n x^n$. The n^{th} term in the series contains the n^{th} power of x. The convergence or divergence of the series depends on the choice of the input variable x.

The *radius of convergence* ρ of $\sum\limits_{n=0}^{\infty} c_n x^n$ can be obtained either by using the n^{th} root test or the ratio test:

$$\frac{1}{\rho} = \lim_{n\to\infty} \sqrt[n]{|c_n|} = \lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right|.$$

The power series $f(x) = \sum\limits_{n=0}^{\infty} c_n x^n$ converges for all $-\rho < x < \rho$.

Explanations

Taylor series

Do you remember your derivative formulas? You can calculate the Taylor series $c_0 + c_1(x-a) + c_2(x-a)^2 + \ldots$ of any function $f(x)$, by choosing c_n equal to the value of the n^{th} derivative of $f(x)$ divided by the appropriate factorial. The more terms you compute in the series, the more accurate your approximation will become.

The zero^th-order approximation to a function $f(x)$ at $x = a$ is

$$f(x) \approx f(a).$$

This approximation is not very accurate in general, but at least it is correct when $x = a$.

The best *linear* approximation to $f(x)$ at $x = a$ is the tangent line $T_1(x)$, which is a line that passes through the point $(a, f(a))$ and has a slope equal to $f'(a)$. Indeed, this is exactly what the first-order Taylor series formula tells us to compute. The coefficient in front of x in the Taylor series is obtained by first calculating $f'(x)$, then evaluating the result at $x = a$:

$$f(x) \approx f(a) + f'(a)(x-a) = T_1(x).$$

To find the best quadratic approximation to $f(x)$, we must compute the second derivative $f''(x)$. The coefficient in front of the x^2 term will be $f''(a)$ divided by $2! = 2$:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 = T_2(x).$$

If we continue like this, we'll obtain the whole Taylor series of the function $f(x)$. At step n, the coefficient c_n will be proportional to the n^{th} derivative of $f(x)$ and the resulting n^{th}-degree polynomial will **imitate the function** in its behaviour up to the n^{th} derivative.

Proof of the sum of the geometric series

We are looking for the sum S_n given by

$$S_n = \sum_{k=0}^{n} r^k = 1 + r + r^2 + r^3 + \cdots + r^n.$$

Observe there is a self-similar pattern in the expanded summation S_n, where subsequent terms gain an additional power of r. Suppose we multiply the above equation by r. This has the effect of "shifting" all the terms to the right:

$$rS_n = r\sum_{k=0}^{n} r^k = r + r^2 + r^3 + \cdots + r^n + r^{n+1}.$$

We can add 1 to both sides to both sides of the equation to obtain

$$1 + rS_n = \underbrace{1 + r + r^2 + r^3 + \cdots + r^n}_{S_n} + r^{n+1} = S_n + r^{n+1}.$$

Note how the sum S_n appears as the first part of the expression on the right-hand side. The resulting equation is $1 + rS_n = S_n + r^{n+1}$. Since we want to find S_n, we can isolate all the S_n terms on one side,

$$1 - r^{n+1} = S_n - rS_n = S_n(1-r),$$

and solve for S_n to obtain $S_n = \frac{1-r^{n+1}}{1-r}$. Neat, huh? This is what math is all about—you spend some time looking for the *structure* in the problem and then exploit this structure to solve the problem using a few lines of arithmetic.

Examples

An infinite series Compute the sum of the infinite series

$$\sum_{n=0}^{\infty} \frac{1}{N+1}\left(\frac{N}{N+1}\right)^n.$$

This may appear complicated, but only until we recognize this is a type of geometric series $\sum ar^n$, where $a = \frac{1}{N+1}$ and $r = \frac{N}{N+1}$:

$$\sum_{n=0}^{\infty} \frac{1}{N+1}\left(\frac{N}{N+1}\right)^n = \sum_{n=0}^{\infty} ar^n = \frac{a}{1-r} = \frac{1}{N+1}\frac{1}{1-\frac{N}{N+1}} = 1.$$

Example Compute $\sin(40°)$ to 15 decimal places. The Maclaurin series of $\sin(x)$ is

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \ldots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}.$$

To calculate the sine of 40 degrees, we compute the sum of the series with x replaced by 40 degrees (expressed in radians). In theory, we need to sum *infinitely* many terms in the series, but in practice we only need to sum the first 8 terms in the series to obtain an accuracy of 15 digits after the decimal. In other words, the series converges very quickly.

Let's use the computer algebra system at `live.sympy.org` to compute the first few terms in the series to see what is going on.

First, we define the n^{th} term:

$$a_n(x) = \frac{(-1)^n x^{2n+1}}{(2n+1)!}.$$

```
>>> def axn_sin(x,n): return (-1.0)**n*x**(2*n+1)/factorial(2*n+1)
```

Next we convert $40°$ to radians:

```
>>> forti = (40*pi/180).evalf()
      0.698131700797732                  # 40 degrees in radians
```

These are the first 10 coefficients in the series:

```
>>> [ axn_sin( forti ,n) for n in range(0,10) ]
 [(0, 0.69813170079773179),              # a_0
```

```
  (1, -0.05671015396488306 2),       # a_1
  (2, 0.0013819920621191727),        # a_2
  (3, -1.6037289757274478e-05),      # a_3
  (4, 1.0856084058295026e-07),       # a_4
  (5, -4.8101124579279279e-10),      # a_5
  (6, 1.5028144059670851e-12),       # a_6
  (7, -3.4878738801065803e-15),      # a_7
  (8, 6.2498067170560129e-18),       # a_8
  (9, -8.9066666494280343e-21)].     # a_9
```

To compute $\sin(40°)$, we sum together all the terms:

```
>>> sum( [ axn_sin( forti ,n) for n in range(0,10) ] )
      0.642787609686539       # the Taylor series approximation

>>> sin(forti).evalf()
      0.642787609686539       # the true value of sin(40)
```

Note the first 8 terms of the series would have been sufficient to obtain an approximation to 15 decimals since the terms a_8 and a_9 are much smaller than 10^{-15}.

Discussion

You can think of the Taylor series as containing the "similarity coefficients" between $f(x)$ and the different powers of x. We choose the terms in the Taylor series of $f(x)$ to ensure the series approximation has the same n^{th} derivative as the function $f(x)$. For a Maclaurin series, the similarity between $f(x)$ and its power series representation is measured at $x = 0$, so the coefficients are chosen as $c_n = \frac{f^{(n)}(0)}{n!}$. The more general Taylor series allows us to build an approximation to $f(x)$ at any point $x = a$, so the similarity coefficients are calculated to match the derivatives at that point: $c_n = \frac{f^{(n)}(a)}{n!}$.

Another way of looking at the Maclaurin series is to imagine it is a kind of X-ray picture for each function $f(x)$. The zero$^{\text{th}}$ coefficient c_0 in the Maclaurin series tells you how much of the constant function is in $f(x)$. The first coefficient,

c_1, tells you how much of the linear function x is in f; the coefficient c_2 tells you about the x^2 contents of f, and so on.

Now get ready for some crazy shit. I want you to go back to page 400 and take a careful look at the Maclaurin series of e^x, $\sin(x)$, and $\cos(x)$. As you will observe, it's as if e^x *contains* both $\sin(x)$ and $\cos(x)$, the only difference being the presence of the alternating negative signs. How about that? Do you remember Euler's formula $e^{ix} = \cos x + i \sin x$? Verify Euler's formula by substituting ix into the power series for e^x.

Another interesting equation to think about in terms of series is $e^x = \cosh x + \sinh x$.

Links

[Animation showing Taylor series approximations to $\sin(x)$]
`http://mathforum.org/mathimages/index.php/Taylor_Series`

5.20 Conclusion

Now you know how to take derivatives, calculate integrals, and find sums of infinite series. These practical skills will come in handy in the future, especially if you choose to pursue a career in science.

The exposure you had to formal math definitions prepared you for more advanced math classes. In particular, you learned how to deal with limits involving infinitely small quantities like ϵ and δ. Recall that both the derivative and the integral are defined as limit expressions. The derivative is defined as a rise-over-run calculation for an infinitely short run. The integral is defined as a Riemann sum with infinitely narrow rectangles.

The notion of a limit is essentially the main new idea we learn in calculus. Limits allow us to talk about infinity. For example, the answer to the question "what is the effective interest rate for a loan with a nominal interest rate of 100%, compounded infinitely often," is equal to the number $e = \lim_{n\to\infty}(1+\frac{1}{n})^n$. If you borrow N dollars today, you will owe Ne dollars at the end of one year. A limit is required to model the "infinitely frequent" compounding.

We also learned some facts about sequences and series. Series teach us how to think about computations with an infinite number of steps. In particular the notion of a Taylor series is a foundational idea for understanding functions.

Above all, the purpose of calculus is to solve problems. Speaking of problems, you should flip the page and dig in. The only way to test if you understand the theory is to apply it in practice.

5.21 Calculus exercises

Calculus hasn't changed much in the last hundred years. The exercises from the book "Calculus Made Easy" by Silvanus Phillips Thompson (originally published[1] in 1910) are equally pertinent and interesting as they were one hundred years ago.

Exercises I. (See p. 437 for Answers.)
Differentiate the following:

(1) $y = x^{13}$ (2) $y = x^{-\frac{3}{2}}$

(3) $y = x^{2a}$ (4) $u = t^{2.4}$

(5) $z = \sqrt[3]{u}$ (6) $y = \sqrt[3]{x^{-5}}$

(7) $u = \sqrt[5]{\dfrac{1}{x^8}}$ (8) $y = 2x^a$

(9) $y = \sqrt[q]{x^3}$ (10) $y = \sqrt[n]{\dfrac{1}{x^m}}$

Exercises II. (See p. 437 for Answers.)
Differentiate the following:

(1) $y = ax^3 + 6.$ (2) $y = 13x^{\frac{3}{2}} - c.$

(3) $y = 12x^{\frac{1}{2}} + c^{\frac{1}{2}}.$ (4) $y = c^{\frac{1}{2}}x^{\frac{1}{2}}.$

(5) $u = \dfrac{az^n - 1}{c}.$ (6) $y = 1.18t^2 + 22.4.$

[1] Full text is available here: `http://bit.ly/tMEtLE` (public domain).

(7) If l_t and l_0 be the lengths of a rod of iron at the temperatures t° C. and 0° C. respectively, then $l_t = l_0(1 + 0.000012t)$. Find the change of length of the rod per degree Centigrade.

(8) It has been found that if c be the candle power of an incandescent electric lamp, and V be the voltage, $c = aV^b$, where a and b are constants.

Find the rate of change of the candle power with the voltage, and calculate the change of candle power per volt at 80, 100 and 120 volts in the case of a lamp for which $a = 0.5 \times 10^{-10}$ and $b = 6$.

(9) The frequency n of vibration of a string of diameter D, length L and specific gravity σ, stretched with a force T, is given by

$$n = \frac{1}{DL}\sqrt{\frac{gT}{\pi\sigma}}.$$

Find the rate of change of the frequency when D, L, σ and T are varied singly.

(10) The greatest external pressure P which a tube can support without collapsing is given by

$$P = \left(\frac{2E}{1-\sigma^2}\right)\frac{t^3}{D^3},$$

where E and σ are constants, t is the thickness of the tube and D is its diameter. (This formula assumes that $4t$ is small compared to D.)

Compare the rate at which P varies for a small change of thickness and for a small change of diameter taking place separately.

(11) Find the rate at which the following vary with respect to a change in radius:

(*a*) the circumference of a circle of radius r;

(*b*) the area of a circle of radius r;

(*c*) the lateral area of a cone of slant dimension l;

(*d*) the volume of a cone of radius r and height h;

(*e*) the area of a sphere of radius r;

(*f*) the volume of a sphere of radius r.

(12) The length L of an iron rod at the temperature T being given by $L = l_t\big[1 + 0.000012(T - t)\big]$, where l_t is the length at the temperature t, find the rate of variation of the diameter D of an iron tyre suitable for being shrunk on a wheel, when the temperature T varies.

Exercises III. (See the Answers on p. 438.)

(1) Differentiate

(*a*) $u = 1 + x + \dfrac{x^2}{1 \cdot 2} + \dfrac{x^3}{1 \cdot 2 \cdot 3} + \cdots$.

(*b*) $y = ax^2 + bx + c$. (*c*) $y = (x + a)^2$.

(*d*) $y = (x + a)^3$.

(2) If $w = at - \frac{1}{2}bt^2$, find $\dfrac{dw}{dt}$.

(3) Find the derivative of

$$y = (x + \sqrt{-1}) \cdot (x - \sqrt{-1}).$$

(4) Differentiate

$$y = (197x - 34x^2) \cdot (7 + 22x - 83x^3).$$

(5) If $x = (y + 3) \cdot (y + 5)$, find $\dfrac{dx}{dy}$.

(6) Differentiate $y = 1.3709x \cdot (112.6 + 45.202x^2)$.

Find the derivative of

(7) $y = \dfrac{2x+3}{3x+2}$.

(8) $y = \dfrac{1+x+2x^2+3x^3}{1+x+2x^2}$.

(9) $y = \dfrac{ax+b}{cx+d}$.

(10) $y = \dfrac{x^n+a}{x^{-n}+b}$.

(11) The temperature t of the filament of an incandescent electric lamp is connected to the current passing through the lamp by the relation

$$C = a + bt + ct^2.$$

Find an expression giving the variation of the current corresponding to a variation of temperature.

(12) The following formulae have been proposed to express the relation between the electric resistance R of a wire at the temperature $t°$C., and the resistance R_0 of that same wire at $0°$ Centigrade, a, b, c being constants.

$$R = R_0(1 + at + bt^2).$$
$$R = R_0(1 + at + b\sqrt{t}).$$
$$R = R_0(1 + at + bt^2)^{-1}.$$

Find the rate of variation of the resistance with regard to temperature as given by each of these formulae.

(13) The electromotive-force E of a certain type of standard cell has been found to vary with the temperature $t[°]$ according to the relation

$$E = 1.4340\bigl[1 - 0.000814(t-15) + 0.000007(t-15)^2\bigr] \text{ volts.}$$

Find the change of electromotive-force per degree, at $15°$, $20°$ and $25°$.

(14) The electromotive-force necessary to maintain an electric arc of length l with a current of intensity i has been found by Mrs. Ayrton to be

$$E = a + bl + \frac{c+kl}{i},$$

where a, b, c, k are constants.

Find an expression for the variation of the electromotive-force (*a*) with regard to the length of the arc; (*b*) with regard to the strength of the current.

Exercises IV. (See page 438 for Answers.)

Find $\dfrac{dy}{dx}$ and $\dfrac{d^2y}{dx^2}$ for the following expressions:

(1) $y = 17x + 12x^2$.

(2) $y = \dfrac{x^2 + a}{x + a}$.

(3) $y = 1 + \dfrac{x}{1} + \dfrac{x^2}{1 \cdot 2} + \dfrac{x^3}{1 \cdot 2 \cdot 3} + \dfrac{x^4}{1 \cdot 2 \cdot 3 \cdot 4}$.

(4) Find the 2nd and 3rd derived functions in the Exercises III. (p. 411), No. 1 to No. 7.

Exercises V. (See page 439 for Answers.)

(1) If $y = a + bt^2 + ct^4$; find $\dfrac{dy}{dt}$ and $\dfrac{d^2y}{dt^2}$.

$$\textit{Ans. } \frac{dy}{dt} = 2bt + 4ct^3; \quad \frac{d^2y}{dt^2} = 2b + 12ct^2.$$

(2) A body falling freely in space describes in t seconds a space s, in feet, expressed by the equation $s = 16t^2$. Draw a curve showing the relation between s and t. Also determine the velocity of the body at the following times from its being let drop: $t = 2$ seconds; $t = 4.6$ seconds; $t = 0.01$ second.

(3) If $x = at - \frac{1}{2}gt^2$; find $\dot{x}$ and $\ddot{x}$.

(4) If a body move according to the law

$$s = 12 - 4.5t + 6.2t^2,$$

find its velocity when $t = 4$ seconds; s being in feet.

(5) Find the acceleration of the body mentioned in the preceding example. Is the acceleration the same for all values of t?

(6) The angle θ (in radians) turned through by a revolving wheel is connected with the time t (in seconds) that has elapsed since starting; by the law

$$\theta = 2.1 - 3.2t + 4.8t^2.$$

Find the angular velocity (in radians per second) of that wheel when $1\frac{1}{2}$ seconds have elapsed. Find also its angular acceleration.

(7) A slider moves so that, during the first part of its motion, its distance s in inches from its starting point is given by the expression

$$s = 6.8t^3 - 10.8t; \quad t \text{ being in seconds.}$$

Find the expression for the velocity and the acceleration at any time; and hence find the velocity and the acceleration after 3 seconds.

(8) The motion of a rising balloon is such that its height h, in miles, is given at any instant by the expression $h = 0.5 + \frac{1}{10}\sqrt[3]{t - 125}$; t being in seconds.

Find an expression for the velocity and the acceleration at any time. Draw curves to show the variation of height, velocity and acceleration during the first ten minutes of the ascent.

(9) A stone is thrown downwards into water and its depth p in metres at any instant t seconds after reaching the surface of the water is given by the expression

$$p = \frac{4}{4 + t^2} + 0.8t - 1.$$

Find an expression for the velocity and the acceleration at any time. Find the velocity and acceleration after 10 seconds.

(10) A body moves in such a way that the spaces described in the time t from starting is given by $s = t^n$, where n is a constant. Find the value of n when the velocity is doubled from the 5th to the 10th second; find it also when the velocity is numerically equal to the acceleration at the end of the 10th second.

Exercises VI. (See page 439 for Answers.)
Differentiate the following:

(1) $y = \sqrt{x^2+1}$.

(2) $y = \sqrt{x^2+a^2}$.

(3) $y = \dfrac{1}{\sqrt{a+x}}$.

(4) $y = \dfrac{a}{\sqrt{a-x^2}}$.

(5) $y = \dfrac{\sqrt{x^2-a^2}}{x^2}$.

(6) $y = \dfrac{\sqrt[3]{x^4+a}}{\sqrt[2]{x^3+a}}$.

(7) $y = \dfrac{a^2+x^2}{(a+x)^2}$.

(8) Differentiate y^5 with respect to y^2.

(9) Differentiate $y = \dfrac{\sqrt{1-\theta^2}}{1-\theta}$.

Exercises VII. You can now successfully try the following. (See page 440 for Answers.)

(1) If $u = \frac{1}{2}x^3$; $v = 3(u+u^2)$; and $w = \dfrac{1}{v^2}$, find $\dfrac{dw}{dx}$.

(2) If $y = 3x^2 + \sqrt{2}$; $z = \sqrt{1+y}$; and $v = \dfrac{1}{\sqrt{3}+4z}$, find $\dfrac{dv}{dx}$.

(3) If $y = \dfrac{x^3}{\sqrt{3}}$; $z = (1+y)^2$; and $u = \dfrac{1}{\sqrt{1+z}}$, find $\dfrac{du}{dx}$.

Exercises VIII. (See page 441 for Answers.)

(1) Plot the curve $y = \frac{3}{4}x^2 - 5$, using a scale of millimetres. Measure at points corresponding to different values of x, the angle of its slope.

Find, by differentiating the equation, the expression for slope; and see, from a Table of Natural Tangents, whether this agrees with the measured angle.

(2) Find what will be the slope of the curve

$$y = 0.12x^3 - 2,$$

at the particular point that has as abscissa $x = 2$.

(3) If $y = (x - a)(x - b)$, show that at the particular point of the curve where $\dfrac{dy}{dx} = 0$, x will have the value $\frac{1}{2}(a + b)$.

(4) Find the $\dfrac{dy}{dx}$ of the equation $y = x^3 + 3x$; and calculate the numerical values of $\dfrac{dy}{dx}$ for the points corresponding to $x = 0$, $x = \frac{1}{2}$, $x = 1$, $x = 2$.

(5) In the curve to which the equation is $x^2 + y^2 = 4$, find the values of x at those points where the slope $= 1$.

(6) Find the slope, at any point, of the curve whose equation is $\dfrac{x^2}{3^2} + \dfrac{y^2}{2^2} = 1$; and give the numerical value of the slope at the place where $x = 0$, and at that where $x = 1$.

(7) The equation of a tangent to the curve $y = 5 - 2x + 0.5x^3$, being of the form $y = mx + n$, where m and n are constants, find the value of m and n if the point where the tangent touches the curve has $x = 2$ for abscissa.

(8) At what angle do the two curves

$$y = 3.5x^2 + 2 \quad \text{and} \quad y = x^2 - 5x + 9.5$$

cut one another?

(9) Tangents to the curve $y = \pm\sqrt{25 - x^2}$ are drawn at points for which $x = 3$ and $x = 4$. Find the coordinates of the point of intersection of the tangents and their mutual inclination.

(10) A straight line $y = 2x - b$ touches a curve $y = 3x^2 + 2$ at one point. What are the coordinates of the point of contact, and what is the value of b?

Exercises IX. (See page 441 for Answers.)

(1) What values of x will make y a maximum and a minimum, if $y = \dfrac{x^2}{x+1}$?

(2) What value of x will make y a maximum in the equation $y = \dfrac{x}{a^2 + x^2}$?

(3) A line of length p is to be cut up into 4 parts and put together as a rectangle. Show that the area of the rectangle will be a maximum if each of its sides is equal to $\frac{1}{4}p$.

(4) A piece of string 30 inches long has its two ends joined together and is stretched by 3 pegs so as to form a triangle. What is the largest triangular area that can be enclosed by the string?

(5) Plot the curve corresponding to the equation

$$y = \frac{10}{x} + \frac{10}{8-x};$$

also find $\dfrac{dy}{dx}$, and deduce the value of x that will make y a minimum; and find that minimum value of y.

(6) If $y = x^5 - 5x$, find what values of x will make y a maximum or a minimum.

(7) What is the smallest square that can be inscribed in a given square?

(8) Inscribe in a given cone, the height of which is equal to the radius of the base, a cylinder (*a*) whose volume is a maximum; (*b*) whose lateral area is a maximum; (*c*) whose total area is a maximum.

(9) Inscribe in a sphere, a cylinder (*a*) whose volume is a maximum; (*b*) whose lateral area is a maximum; (*c*) whose total area is a maximum.

(10) A spherical balloon is increasing in volume. If, when its radius is r feet, its volume is increasing at the rate of 4 cubic feet per second, at what rate is its surface then increasing?

(11) Inscribe in a given sphere a cone whose volume is a maximum.

(12) The current C given by a battery of N similar voltaic cells is $C = \dfrac{n \cdot E}{R + \dfrac{rn^2}{N}}$, where E, R, r, are constants and n is the number of cells coupled in series. Find the proportion of n to N for which the current is greatest.

Exercises X. (You are advised to plot the graph of any numerical example.) (See p. 442 for the Answers.)

(1) Find the maxima and minima of

$$y = x^3 + x^2 - 10x + 8.$$

(2) Given $y = \dfrac{b}{a}x - cx^2$, find expressions for $\dfrac{dy}{dx}$, and for $\dfrac{d^2y}{dx^2}$, also find the value of x which makes y a maximum or a minimum, and show whether it is maximum or minimum.

(3) Find how many maxima and how many minima there are in the curve, the equation to which is

$$y = 1 - \frac{x^2}{2} + \frac{x^4}{24};$$

and how many in that of which the equation is

$$y = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720}.$$

(4) Find the maxima and minima of

$$y = 2x + 1 + \frac{5}{x^2}.$$

(5) Find the maxima and minima of

$$y = \frac{3}{x^2 + x + 1}.$$

(6) Find the maxima and minima of

$$y = \frac{5x}{2 + x^2}.$$

(7) Find the maxima and minima of

$$y = \frac{3x}{x^2 - 3} + \frac{x}{2} + 5.$$

(8) Divide a number N into two parts in such a way that three times the square of one part plus twice the square of the other part shall be a minimum.

(9) The efficiency u of an electric generator at different values of output x is expressed by the general equation:

$$u = \frac{x}{a + bx + cx^2};$$

where a is a constant depending chiefly on the energy losses in the iron and c a constant depending chiefly on the resistance of the copper parts. Find an expression for that value of the output at which the efficiency will be a maximum.

(10) Suppose it to be known that consumption of coal by a certain steamer may be represented by the formula $y = 0.3 + 0.001v^3$; where y is the number of tons of coal burned per hour and v is the speed expressed in nautical miles per hour. The cost of wages, interest on capital, and depreciation of that ship are together equal, per hour, to the cost of 1 ton of coal. What speed will make the total cost of a voyage of 1000 nautical miles a minimum? And, if coal costs 10 shillings per ton, what will that minimum cost of the voyage amount to?

(11) Find the maxima and minima of

$$y = \pm \frac{x}{6}\sqrt{x(10 - x)}.$$

(12) Find the maxima and minima of

$$y = 4x^3 - x^2 - 2x + 1.$$

Exercises XI. (See page 442 for Answers.)
Split into fractions:

(1) $\dfrac{3x+5}{(x-3)(x+4)}$.

(2) $\dfrac{3x-4}{(x-1)(x-2)}$.

(3) $\dfrac{3x+5}{x^2+x-12}$.

(4) $\dfrac{x+1}{x^2-7x+12}$.

(5) $\dfrac{x-8}{(2x+3)(3x-2)}$.

(6) $\dfrac{x^2-13x+26}{(x-2)(x-3)(x-4)}$.

(7) $\dfrac{x^2-3x+1}{(x-1)(x+2)(x-3)}$.

(8) $\dfrac{5x^2+7x+1}{(2x+1)(3x-2)(3x+1)}$.

(9) $\dfrac{x^2}{x^3-1}$.

(10) $\dfrac{x^4+1}{x^3+1}$.

(11) $\dfrac{5x^2+6x+4}{(x+1)(x^2+x+1)}$.

(12) $\dfrac{x}{(x-1)(x-2)^2}$.

(13) $\dfrac{x}{(x^2-1)(x+1)}$.

(14) $\dfrac{x+3}{(x+2)^2(x-1)}$.

(15) $\dfrac{3x^2+2x+1}{(x+2)(x^2+x+1)^2}$.

(16) $\dfrac{5x^2+8x-12}{(x+4)^3}$.

(17) $\dfrac{7x^2+9x-1}{(3x-2)^4}$.

(18) $\dfrac{x^2}{(x^3-8)(x-2)}$.

Exercises XII. (See page 443 for Answers.)

(1) Differentiate $y = b(e^{ax} - e^{-ax})$.

(2) Find the derivative with respect to t of $u(t) = at^2 + 2\ln t$.

(3) If $y = n^t$, find $\dfrac{d(\ln y)}{dt}$.

(4) Show that if $y = \dfrac{1}{b} \cdot \dfrac{a^{bx}}{\ln a}$, $\dfrac{dy}{dx} = a^{bx}$.

(5) If $w = pv^n$, find $\dfrac{dw}{dv}$.

Differentiate

(6) $y = \ln x^n$.

(7) $y = 3e^{-\frac{x}{x-1}}$.

(8) $y = (3x^2 + 1)e^{-5x}$.

(9) $y = \ln(x^a + a)$.

(10) $y = (3x^2 - 1)(\sqrt{x} + 1)$.

(11) $y = \dfrac{\ln(x+3)}{x+3}$.

(12) $y = a^x \cdot x^a$.

(13) It was shown by Lord Kelvin that the speed of signalling through a submarine cable depends on the value of the ratio of the external diameter of the core to the diameter of the enclosed copper wire. If this ratio is called y, then the number of signals s that can be sent per minute can be expressed by the formula

$$s = ay^2 \ln \frac{1}{y};$$

where a is a constant depending on the length and the quality of the materials. Show that if these are given, s will be a maximum if $y = 1/\sqrt{e}$.

(14) Find the maximum or minimum of

$$y = x^3 - \ln x.$$

(15) Differentiate $y = \ln(axe^x)$.

(16) Differentiate $y = (\ln ax)^3$.

Exercises XIII. (See page 444 for Answers.)

(1) Draw the curve $y = be^{-\frac{t}{T}}$; where $b = 12$, $T = 8$, and t is given various values from 0 to 20.

(2) If a hot body cools so that in 24 minutes its excess of temperature has fallen to half the initial amount, deduce the time-constant, and find how long it will be in cooling down to 1 per cent. of the original excess.

(3) Plot the curve $y = 100(1 - e^{-2t})$.

(4) The following equations give very similar curves:

$$\text{(i) } y = \frac{ax}{x+b};$$
$$\text{(ii) } y = a(1 - e^{-\frac{x}{b}});$$
$$\text{(iii) } y = \frac{a}{90^\circ} \tan^{-1}\left(\frac{x}{b}\right).$$

Draw all three curves, taking $a = 100$ millimetres; $b = 30$ millimetres.

(5) Find the derivative of y with respect to x, if

$$(a)\ y = x^x; \quad (b)\ y = (e^x)^x; \quad (c)\ y = e^{x^x}.$$

(6) For "Thorium A," the value of λ is 5; find the "mean life," that is, the time taken by the transformation of a quantity Q of "Thorium A" equal to half the initial quantity Q_0 in the expression

$$Q = Q_0 e^{-\lambda t};$$

t being in seconds.

(7) A condenser of capacitance $C = 4.0 \times 10^{-6}$, charged to a potential $V_0 = 20$, is discharging through a resistance of $R = 10,000$ ohms. Find the potential V after (*a*) 0.1 second; (*b*) 0.01 second; assuming that the fall of potential follows the rule $V = V_0 e^{-\frac{t}{RC}}$.

(8) The charge Q of an electrified insulated metal sphere is reduced from 20 to 16 units in 10 minutes. Find the coefficient μ of leakage, if $Q = Q_0 \cdot e^{-\mu t}$; Q_0 being the initial charge and t being in seconds. Hence find the time taken by half the charge to leak away.

(9) The damping on a telephone line can be ascertained from the relation $i = i_0 e^{-\beta l}$, where i is the strength, after t seconds, of a telephonic current of initial strength i_0; l is the length of the line in kilometres, and β is a constant. For the Franco-English submarine cable laid in 1910, $\beta = 0.0114$. Find the damping at the end of the cable (40 kilometres), and the length along which i is still 8% of the original current (limiting value of very good audition).

(10) The pressure p of the atmosphere at an altitude h kilometres is given by $p = p_0 e^{-kh}$; p_0 being the pressure at sea-level (760 millimetres).

The pressures at 10, 20 and 50 kilometres being 199.2, 42.2, 0.32 respectively, find k in each case. Using the mean value of k, find the percentage error in each case.

(11) Find the minimum or maximum of $y = x^x$.

(12) Find the minimum or maximum of $y = x^{\frac{1}{x}}$.

(13) Find the minimum or maximum of $y = xa^{\frac{1}{x}}$.

Exercises XIV. (See page 445 for Answers.)

(1) Differentiate the following:

$$\text{(i)} \quad y = A \sin\left(\theta - \frac{\pi}{2}\right).$$

$$\text{(ii)} \quad y = \sin^2\theta; \quad \text{and } y = \sin 2\theta.$$

$$\text{(iii)} \quad y = \sin^3\theta; \quad \text{and } y = \sin 3\theta.$$

(2) Find the value of θ for which $\sin\theta \cdot \cos\theta$ is a maximum.

(3) Differentiate $y = \dfrac{1}{2\pi}\cos 2\pi nt$.

(4) If $y = \sin a^x$, find $\dfrac{dy}{dx}$.

(5) Differentiate $y = \ln\cos x$.

(6) Differentiate $y = 18.2\sin(x + 26^\circ)$.

(7) Plot the curve $y = 100\sin(\theta - 15^\circ)$; and show that the slope of the curve at $\theta = 75^\circ$ is half the maximum slope.

(8) If $y = \sin\theta \cdot \sin 2\theta$, find $\dfrac{dy}{d\theta}$.

(9) If $y = a \cdot \tan^m(\theta^n)$, find the derivative of y with respect to θ.

(10) Differentiate $y = e^x \sin^2 x$.

(11) Differentiate the three equations of Exercises XIII. (p. 422), No. 4, and compare their derivatives, as to whether they are equal, or nearly equal, for very small values of x, or for very large values of x, or for values of x in the neighbourhood of $x = 30$.

(12) Differentiate the following:

(i) $y = \sec x$.

(ii) $y = \cos^{-1}(x)$.

(iii) $y = \tan^{-1}(x)$.

(iv) $y = \sec^{-1}(x)$.

(v) $y = \tan x \cdot \sqrt{3\sec x}$.

(13) Differentiate $y = \sin(2\theta + 3)^{2.3}$.

(14) Differentiate $y = \theta^3 + 3\sin(\theta + 3) - 3^{\sin\theta} - 3^\theta$.

(15) Find the maximum or minimum of $y = \theta\cos\theta$.

Exercises XVI. (See page 446 for Answers.)

(1) Find the sum of $\frac{2}{3}+\frac{1}{3}+\frac{1}{6}+\frac{1}{12}+\frac{1}{24}+$ etc.

(2) Show that the series $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\frac{1}{7}$ etc., is convergent, and find its sum to 8 terms.

(3) If $\ln(1+x)=x-\dfrac{x^2}{2}+\dfrac{x^3}{3}-\dfrac{x^4}{4}+$ etc., find $\ln 1.3$.

(4) Following a reasoning similar to that explained in this chapter, find y,

$$(a) \text{ if } \frac{dy}{dx}=\tfrac{1}{4}x; \quad (b) \text{ if } \frac{dy}{dx}=\cos x.$$

(5) If $\dfrac{dy}{dx}=2x+3$, find y.

Exercises XVII. (See p. 446 for the Answers.)

(1) Find $\displaystyle\int y\,dx$ when $y^2=4ax$.

(2) Find $\displaystyle\int \frac{3}{x^4}\,dx$.

(3) Find $\displaystyle\int \frac{1}{a}x^3\,dx$.

(4) Find $\displaystyle\int (x^2+a)\,dx$.

(5) Integrate $5x^{-\frac{7}{2}}$.

(6) Find $\displaystyle\int (4x^3+3x^2+2x+1)\,dx$.

(7) If $\dfrac{dy}{dx}=\dfrac{ax}{2}+\dfrac{bx^2}{3}+\dfrac{cx^3}{4}$; find y.

(8) Find $\displaystyle\int \left(\frac{x^2+a}{x+a}\right)dx$.

(9) Find $\displaystyle\int (x+3)^3\,dx$.

(10) Find $\displaystyle\int (x+2)(x-a)\,dx$.

(11) Find $\int (\sqrt{x} + \sqrt[3]{x}) 3a^2\, dx$.

(12) Find $\int (\sin\theta - \tfrac{1}{2}) \frac{d\theta}{3}$.

(13) Find $\int \cos^2 a\theta\, d\theta$.

(14) Find $\int \sin^2 \theta\, d\theta$.

(15) Find $\int \sin^2 a\theta\, d\theta$.

(16) Find $\int \epsilon^{3x}\, dx$.

(17) Find $\int \frac{1}{1+x}\, dx$.

(18) Find $\int \frac{}{1-x}\, dx$.

Exercises XVIII. (See p. 447 for Answers.)

(1) Find the area of the curve $y = x^2 + x - 5$ between $x = 0$ and $x = 6$, and the mean ordinates between these limits.

(2) Find the area of the parabola $y = 2a\sqrt{x}$ between $x = 0$ and $x = a$. Show that it is two-thirds of the rectangle of the limiting ordinate and of its abscissa.

(3) Find the area of the positive portion of a sine curve and the mean ordinate.

(4) Find the area of the positive portion of the curve $y = \sin^2 x$, and find the mean ordinate.

(5) Find the area included between the two branches of the curve $y = x^2 \pm x^{\frac{5}{2}}$ from $x = 0$ to $x = 1$, also the area of the positive portion of the lower branch of the curve.

(6) Find the volume of a cone of radius of base r, and of height h.

(7) Find the area of the curve $y = x^3 - \ln x$ between $x = 0$ and $x = 1$.

(8) Find the volume generated by the curve $y = \sqrt{1 + x^2}$, as it revolves about the axis of x, between $x = 0$ and $x = 4$.

(9) Find the volume generated by a sine curve revolving about the axis of x. Find also the area of its surface.

(10) Find the area of the portion of the curve $xy = a$ included between $x = 1$ and $x = a$. Find the mean ordinate between these limits.

(11) Show that the quadratic mean of the function $y = \sin x$, between the limits of 0 and π radians, is $\dfrac{\sqrt{2}}{2}$. Find also the arithmetical mean of the same function between the same limits; and show that the form-factor is $= 1.11$.

(12) Find the arithmetical and quadratic means of the function $x^2 + 3x + 2$, from $x = 0$ to $x = 3$.

(13) Find the quadratic mean and the arithmetical mean of the function $y = A_1 \sin x + A_1 \sin 3x$.

(14) A certain curve has the equation $y = 3.42e^{0.21x}$. Find the area included between the curve and the axis of x, from the ordinate at $x = 2$ to the ordinate at $x = 8$. Find also the height of the mean ordinate of the curve between these points.

(15) Show that the radius of a circle, the area of which is twice the area of a polar diagram, is equal to the quadratic mean of all the values of r for that polar diagram.

(16) Find the volume generated by the curve $y = \pm\dfrac{x}{6}\sqrt{x(10-x)}$ rotating about the axis of x.

Exercises XIX. (See p. 449 for Answers.)

(1) Find $\displaystyle\int \sqrt{a^2 - x^2}\, dx$.

(2) Find $\displaystyle\int x \ln x\, dx$.

(3) Find $\displaystyle\int x^a \ln x\, dx$.

(4) Find $\displaystyle\int e^x \cos e^x\, dx$.

(5) Find $\displaystyle\int \frac{1}{x}\cos(\ln x)\, dx$.

(6) Find $\displaystyle\int x^2 e^x\, dx$.

(7) Find $\displaystyle\int \frac{(\ln x)^a}{x}\,dx$.

(8) Find $\displaystyle\int \frac{1}{x\ln x}\,dx$.

(9) Find $\displaystyle\int \frac{5x+1}{x^2+x-2}\,dx$.

(10) Find $\displaystyle\int \frac{(x^2-3)}{x^3-7x+6}\,dx$.

(11) Find $\displaystyle\int \frac{b}{x^2-a^2}\,dx$.

(12) Find $\displaystyle\int \frac{4x}{x^4-1}\,dx$.

(13) Find $\displaystyle\int \frac{1}{1-x^4}\,dx$.

(14) Find $\displaystyle\int \frac{1}{x\sqrt{a-bx^2}}\,dx$.

End matter

Conclusion

We managed to cover a lot of ground, explaining many topics and concepts in a relatively small textbook. We reviewed high school math and learned about mechanics and calculus. Above all, we examined math and physics material in an integrated manner.

If you liked or hated this book, be sure to send me feedback. Feedback is crucial so I know how to adjust the writing, the content, and the attitude of the book for future learners of math. Please take the time to drop me a line and let me know what you thought. My email address is ivan.savov@gmail.com.

I have a followup physics book on electricity and magnetism in the works. Another title on linear algebra is also nearly complete. To stay informed about new titles follow me on Twitter @minireference or check out the company blog at http://minireference.com/blog/.

Acknowledgments

This book would not have been possible without the support and encouragement of the people around me. I am fortunate to have grown up surrounded by good people who knew the value of math and encouraged me in my studies and with this project. In this section, I want to *big up* all the people who deserve it.

First and foremost in this list are my parents from whom I have learned many things, and who have supported me throughout my life.

Next in line are all my teachers. I thank my CEGEP teachers: Karnig Bedrossian from whom I learned calculus, Paul Kenton from whom I learned how to think about physics in a chill manner, and Benoit Larose who taught me that more dimensions does not mean things get more complicated. I thank Kohur Gowrisankaran, Frank Ferrie, Mourad El-Gamal, and Ioannis Psaromiligkos for their teaching during my engineering days, and Guy Moore and Zaven Altounian for teaching me advanced physics topics. Among all my teachers, I owe the most to Patrick Hayden whose teaching methods have always inspired me. From him, I learned that by defining things clearly, you can *trick* students into learning advanced topics, and even make it seem that the results are obvious! Thanks go out to all my research collaborators and friends: David Avis, Arlo Breault, Juan Pablo Di Lelle, Omar Fawzi, Adriano Ferrari, Igor Khavkine, Doina Precup, Andie Sigler, and Mark M. Wilde. Thank you all for teaching me a great many things!

Preparing this book took a lot of effort. I want to thank Afton Lewis, Oleg Zhoglo, and Alexandra Foty for helping me proofread v2.0 of the book, and all the readers of v3.0 who suggested clarifications. Thank you all for your comments and feedback. Above all, I want to thank my editor Sandy Gordon, who helped me prepare v4.0 of the book, which is a substantial improvement over previous versions. Her expertise with the English language and her advice on style and content have been invaluable.

Last but not least, I want to thank all my students for their endless questions and demands for explanations. If I have developed any skill for explaining things, I owe it to them.

Further reading

You have reached the end of this book, but you are only at the beginning of the journey of scientific discovery. There are a lot of cool things left for you to learn about. Below are some recommendation of subjects you might find interesting.

Electricity and Magnetism

Electrostatics is the study of the electric force $\vec{F}_e$ and the associated electric potential U_e. Here, you will also learn about the electric field $\vec{E}$ and electric potential V.

Magnetism is the study of the magnetic force $\vec{F}_b$ and the magnetic field $\vec{B}$, which are caused by electric currents flowing through wires. The current I is the total number of electrons passing through a cross-section of the wire in one second. By virtue of its motion through space, each electron contributes to the strength of the magnetic field surrounding the wire.

The beauty of electromagnetism is that the entire theory can be described in just four equations:

$$\begin{aligned} \nabla \cdot \vec{E} &= \frac{\rho}{\varepsilon_0} && \text{Gauss's law} \\ \nabla \cdot \vec{B} &= 0 && \text{Gauss's law for magnetism} \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} && \text{Faraday's law of induction} \\ \nabla \times \vec{B} &= \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t} && \text{Ampère's circuital law} \end{aligned}$$

Together, these are known as Maxwell's equations.

Vector calculus

You may be wondering what the triangle thing ∇ is. The symbol ∇ (nabla) is the vector derivative operation. Guess what—you can also do calculus with vectors.

In vector calculus you will learn about path integrals, surface integrals, and volume integrals of vector quantities. You will also learn about vector-derivatives, as well as two vector equivalents of the fundamental theorem of calculus:

- Stokes' Theorem:

$$\iint_{\Sigma} \nabla \times \vec{F} \cdot d\vec{\Sigma} = \int_{\partial\Sigma} \vec{F} \cdot d\vec{r},$$

which relates the integral of the $\mathrm{curl}\vec{F} \equiv \nabla \times \vec{F}$ of the field $\vec{F}$ over the surface Σ to the circulation of $\vec{F}$ along the boundary of the surface $\partial\Sigma$.

- Gauss' Divergence Theorem:

$$\iiint_{\mathrm{V}} \nabla \cdot \vec{F}\; d\mathrm{V} = \iint_{\partial\mathrm{V}} \vec{F} \cdot d\vec{\Sigma},$$

which relates the integral of the divergence $\mathrm{div}\vec{F} \equiv \nabla \cdot \vec{F}$ of the field $\vec{F}$ over the volume V to the flux of $\vec{F}$ through the volume boundary $\partial\mathrm{V}$.

Both of these theorems relate the total of some derivative quantity over some region R to the quantity on the boundary of the region R, which we denote as ∂R. The fundamental theorem of calculus can also be interpreted in the same manner:

$$\int_I F'(x)\, dx = \int_a^b F'(x)\, dx = F_{\partial I} = F(b) - F(a),$$

where $I = [a, b]$ is the *interval* from a to b on the real line and the two points a and b form its boundary ∂I.

This course will be of interest mainly for physicists and engineers.

Multivariable calculus

Of wider interest is the study of calculus with functions that have more than one input variable. Consider as an example a function $f(x, y)$ that has two input variables, x and y. You can plot this function as a *surface*, where the height z of the surface above the point (x, y) is given by the function value $z = f(x, y)$.

There is no new math to learn in multivariable calculus: it's the same stuff as differential calculus and integral calculus but with more variables. For a function $f(x, y)$, there will be an "x-derivative" $\frac{\partial}{\partial x}$ and a "y-derivative" $\frac{\partial}{\partial y}$. The operator ∇ is a combination of both the x and y derivatives: $\nabla f(x, y) = [\frac{\partial f}{\partial x}, \frac{\partial f}{\partial x}]$. Note that ∇ acts on a function $f(x, y)$ to produce a vector. This is known as the *gradient* vector, which tells you the "slope" of the surface. More specifically, the gradient vector tells you the direction of the function's maximum increase. If you think of $z = f(x, y)$ as the height of a mountain at particular (x, y) coordinates on a map, then the gradient vector $\nabla f(x, y)$ always points uphill.

If you understand derivatives and integrals, then you will find this course very easy.

Probability

Probability distributions are a fundamental tool for modelling non-deterministic behaviour. A discrete random variable X is associated with a probability mass function $p_X(x) \equiv \Pr\{X = x\}$, which assigns a "probability mass" to each of the possible outcomes $x \in \mathcal{X}$. For example, if X represents the outcome of the throw of a fair die, then the possible outcomes are $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and the probability mass function has the values $p_X(x) = \frac{1}{6}$, $\forall x \in \mathcal{X}$.

Probability theory is used all over the place: statistics, machine learning, quantum mechanics, gambling, risk analysis, etc.

General mathematics

Mathematics is a very broad field. There are all kinds of topics to learn about; some of them fun, some of them useful, some of them boring, and some of them which are mind expanding.

I recently discovered a book that covers many math topics of general interest and serves as a great overview of the many areas of mathematics. I highly recommend you take a look at this book.

[BOOK] Richard Elwes. *Mathematics 1001: Absolutely Everything That Matters About Mathematics in 1001 Bite-Sized Explanations*, Firefly Books, 2010, ISBN 1554077192.

General physics

If you want to learn more about physics, I highly recommend the Feynman lectures on physics. This three-tome collection covers all of undergraduate physics with countless links to more advanced topics. While on the Feynman note, I want to also recommend his other book with life stories.

[BOOK] Richard P. Feynman, Robert B. Leighton, Matthew Sands. *The Feynman Lectures on Physics including Feynman's Tips on Physics: The Definitive and Extended Edition*, Addison Wesley, 2005, ISBN 0805390456.
[BOOK] Richard P. Feynman. *Surely You're Joking, Mr. Feynman! (Adventures of a Curious Character)*, W. W. Norton & Company, 1997, ISBN 0393316041.

Lagrangian mechanics

In this book we learned about *Newtonian mechanics*, that is, mechanics starting from Newton's laws. There is a much more general framework known as Lagrangian mechanics which can be used to analyze more complex mechanical systems. The following is an excellent book on the subject.
[BOOK] Herbert Goldstein, Charles P. Poole Jr., John L. Safko. *Classical Mechanics*, Addison-Wesley, Third edition, 2001, ISBN 0201657023.

Quantum mechanics

Quantum mechanics describes the physics of all things is small: photons, electrons and atoms. An absolutely approachable and readable introduction to the subject is Richard Feynman's QED book.

For a deeper understanding of quantum mechanics, I recommend the book by Sakurai. If you understand linear algebra, then you can understand quantum mechanics.

[BOOK] Richard P. Feynman. *QED: The strange theory of light and matter*. Princeton University Press, 2006, ISBN 0691125759.
[BOOK] Jun John Sakurai. *Modern Quantum Mechanics*, Second Edition, Addison-Wesley, 2010, ISBN 0805382917.

Information theory

Claude Shannon developed a mathematical framework for studying the problems of information storage and information transmission. Using statistical notions such as entropy, we can quantify the information content of data sources and the information transmitting abilities of noisy communication channels.

We can arrive at an *operational* interpretation of the information carrying capacity of a noisy communication channel in terms of our ability to convert it into a noiseless channel. Channels with more noise have a smaller capacity for carrying information. Consider a channel that allows us to send data at the rate of 1[MB/sec], on which half of the packets sent get lost due to the effects of noise on the channel. It is not true that the capacity of such a channel is 1[MB/sec], because we must also account for the need to retransmit lost packets. To correctly characterize a channel's information carrying capacity, we must consider the rate of the end-to-end *code*, which converts many uses of the noisy channel into an effectively noiseless communication channel.

Channel coding is one of the fundamental problems studied in information theory. I highly recommend the excellent textbook on the subject by Cover and Thomas.

[BOOK] Thomas M. Cover, Joy A. Thomas. *Elements of Information Theory*, Wiley, 2006, ISBN 0471241954.

Final words

Throughout this book, I strived to equip you with the tools you'll need to make your future science studies enjoyable and pain free. Remember to always take it easy. Play with math and never take things too seriously. Grades don't matter. Big paycheques don't matter. Never settle for a boring job just because it pays well. Try to work only on projects you care about.

I want you to be confident in your ability to handle math, physics, and the other complicated stuff life will throw at you. You have the tools to do anything you want; choose your own adventure. And if the big banks come-a-knocking one day with a big paycheque trying to bribe you into applying your analytical skills to their avaricious schemes, you can send them-a-walking.

Answers to exercises

Exercises I. (p. 409.)

(1) $\dfrac{dy}{dx} = 13x^{12}$. (2) $\dfrac{dy}{dx} = -\dfrac{3}{2}x^{-\frac{5}{2}}$. (3) $\dfrac{dy}{dx} = 2ax^{(2a-1)}$.

(4) $\dfrac{du}{dt} = 2.4t^{1.4}$. (5) $\dfrac{dz}{du} = \dfrac{1}{3}u^{-\frac{2}{3}}$. (6) $\dfrac{dy}{dx} = -\dfrac{5}{3}x^{-\frac{8}{3}}$.

(7) $\dfrac{du}{dx} = -\dfrac{8}{5}x^{-\frac{13}{5}}$. (8) $\dfrac{dy}{dx} = 2ax^{a-1}$.

(9) $\dfrac{dy}{dx} = \dfrac{3}{q}x^{\frac{3-q}{q}}$. (10) $\dfrac{dy}{dx} = -\dfrac{m}{n}x^{-\frac{m+n}{n}}$.

Exercises II. (p. 409.)

(1) $\dfrac{dy}{dx} = 3ax^2$. (2) $\dfrac{dy}{dx} = 13 \times \frac{3}{2}x^{\frac{1}{2}}$. (3) $\dfrac{dy}{dx} = 6x^{-\frac{1}{2}}$.

(4) $\dfrac{dy}{dx} = \dfrac{1}{2}c^{\frac{1}{2}}x^{-\frac{1}{2}}$. (5) $\dfrac{du}{dz} = \dfrac{an}{c}z^{n-1}$. (6) $\dfrac{dy}{dt} = 2.36t$.

(7) $\dfrac{dl_t}{dt} = 0.000012 \times l_0$.

(8) $\dfrac{dC}{dV} = abV^{b-1}$, 0.98, 3.00 and 7.47 candle power per volt respectively.

(9) $$\frac{dn}{dD} = -\frac{1}{LD^2}\sqrt{\frac{gT}{\pi\sigma}}, \quad \frac{dn}{dL} = -\frac{1}{DL^2}\sqrt{\frac{gT}{\pi\sigma}},$$
$$\frac{dn}{d\sigma} = -\frac{1}{2DL}\sqrt{\frac{gT}{\pi\sigma^3}}, \quad \frac{dn}{dT} = \frac{1}{2DL}\sqrt{\frac{g}{\pi\sigma T}}.$$

(10) $$\frac{\text{Rate of change of } P \text{ when } t \text{ varies}}{\text{Rate of change of } P \text{ when } D \text{ varies}} = -\frac{D}{t}.$$

(11) 2π, $2\pi r$, πl, $\frac{2}{3}\pi rh$, $8\pi r$, $4\pi r^2$. (12) $\dfrac{dD}{dT} = \dfrac{0.000012l_t}{\pi}$.

Exercises III. (p. 411.)

(1) (*a*) $1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$ (*b*) $2ax + b.$ (*c*) $2x + 2a.$
(*d*) $3x^2 + 6ax + 3a^2.$

(2) $\frac{dw}{dt} = a - bt.$ (3) $\frac{dy}{dx} = 2x.$

(4) $14110x^4 - 65404x^3 - 2244x^2 + 8192x + 1379.$

(5) $\frac{dx}{dy} = 2y + 8.$ (6) $185.9022654x^2 + 154.36334.$

(7) $\frac{-5}{(3x+2)^2}.$ (8) $\frac{6x^4 + 6x^3 + 9x^2}{(1 + x + 2x^2)^2}.$

(9) $\frac{ad - bc}{(cx + d)^2}.$ (10) $\frac{anx^{-n-1} + bnx^{n-1} + 2nx^{-1}}{(x^{-n} + b)^2}.$

(11) $b + 2ct.$

(12) $R_0(a + 2bt), \quad R_0\left(a + \frac{b}{2\sqrt{t}}\right), \quad -\frac{R_0(a + 2bt)}{(1 + at + bt^2)^2}$ or $\frac{R^2(a + 2bt)}{R_0}.$

(13) $1.4340(0.000014t - 0.001024), \quad -0.00117, \quad -0.00107, \quad -0.00097.$

(14) $\frac{dE}{dl} = b + \frac{k}{i}, \quad \frac{dE}{di} = -\frac{c + kl}{i^2}.$

Exercises IV. (p. 413.)

(1) $17 + 24x; \quad 24.$ (2) $\frac{x^2 + 2ax - a}{(x + a)^2}; \quad \frac{2a(a + 1)}{(x + a)^3}.$

(3) $1 + x + \frac{x^2}{1 \times 2} + \frac{x^3}{1 \times 2 \times 3}; \quad 1 + x + \frac{x^2}{1 \times 2}.$

(4) (*Exercises III.*):

(1) (*a*) $\dfrac{d^2y}{dx^2} = \dfrac{d^3y}{dx^3} = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \dots$.
(*b*) $2a$, 0. (*c*) 2, 0. (*d*) $6x + 6a$, 6.

(2) $-b$, 0. (3) 2, 0.

(4) $56440x^3 - 196212x^2 - 4488x + 8192$.

$169320x^2 - 392424x - 4488$.

(5) 2, 0. (6) $371.80453x$, 371.80453.

(7) $\dfrac{30}{(3x+2)^3}$, $-\dfrac{270}{(3x+2)^4}$.

Exercises V. (p. 413.)

(2) 64; 147.2; and 0.32 feet per second.

(3) $x = a - gt$; $\ddot{x} = -g$. (4) 45.1 feet per second.

(5) 12.4 feet per second per second. Yes.

(6) Angular velocity $= 11.2$ radians per second; angular acceleration $= 9.6$ radians per second per second.

(7) $v = 20.4t^2 - 10.8$. $a = 40.8t$. 172.8 in./sec., 122.4 in./sec^2.

(8) $v = \dfrac{1}{30\sqrt[3]{(t-125)^2}}$, $a = -\dfrac{1}{45\sqrt[3]{(t-125)^5}}$.

(9) $v = 0.8 - \dfrac{8t}{(4+t^2)^2}$, $a = \dfrac{24t^2 - 32}{(4+t^2)^3}$, 0.7926 and 0.00211.

(10) $n = 2$, $n = 11$.

Exercises VI. (p. 415.)

(1) $\dfrac{x}{\sqrt{x^2+1}}$.

(2) $\dfrac{x}{\sqrt{x^2+a^2}}$.

(3) $-\dfrac{1}{2\sqrt{(a+x)^3}}$.

(4) $\dfrac{ax}{\sqrt{(a-x^2)^3}}$.

(5) $\dfrac{2a^2-x^2}{x^3\sqrt{x^2-a^2}}$.

(6) $\dfrac{\frac{3}{2}x^2\left[\frac{8}{9}x\left(x^3+a\right)-\left(x^4+a\right)\right]}{(x^4+a)^{\frac{2}{3}}(x^3+a)^{\frac{3}{2}}}$

(7) $\dfrac{2a\left(x-a\right)}{(x+a)^3}$.

(8) $\frac{5}{2}y^3$.

(9) $\dfrac{1}{(1-\theta)\sqrt{1-\theta^2}}$.

Exercises VII. (p. 415.)

(1) $\dfrac{dw}{dx}=\dfrac{3x^2\left(3+3x^3\right)}{27\left(\frac{1}{2}x^3+\frac{1}{4}x^6\right)^3}$.

(2) $\dfrac{dv}{dx}=-\dfrac{12x}{\sqrt{1+\sqrt{2}+3x^2}\left(\sqrt{3}+4\sqrt{1+\sqrt{2}+3x^2}\right)^2}$.

(3) $\dfrac{du}{dx}=-\dfrac{x^2\left(\sqrt{3}+x^3\right)}{\sqrt{\left[1+\left(1+\dfrac{x^3}{\sqrt{3}}\right)^2\right]^3}}$

Exercises VIII. (p. 415.)

(2) 1.44.

(4) $\dfrac{dy}{dx} = 3x^2 + 3$; and the numerical values are: 3, $3\frac{3}{4}$, 6, and 15.

(5) $\pm\sqrt{2}$.

(6) $\dfrac{dy}{dx} = -\dfrac{4}{9}\dfrac{x}{y}$. Slope is zero where $x = 0$; and is $\mp\dfrac{1}{3\sqrt{2}}$ where $x = 1$.

(7) $m = 4$, $n = -3$.

(8) Intersections at $x = 1$, $x = -3$. Angles $153^\circ 26'$, $2^\circ 28'$.

(9) Intersection at $x = 3.57$, $y = 3.50$. Angle $16^\circ 16'$.

(10) $x = \frac{1}{3}$, $y = 2\frac{1}{3}$, $b = -\frac{5}{3}$.

Exercises IX. (p. 417.)

(1) Min.: $x = 0$, $y = 0$; max.: $x = -2$, $y = -4$.

(2) $x = a$.

(4) $25\sqrt{3}$ square inches.

(5) $\dfrac{dy}{dx} = -\dfrac{10}{x^2} + \dfrac{10}{(8-x)^2}$; $x = 4$; $y = 5$.

(6) Max. for $x = -1$; min. for $x = 1$.

(7) Join the middle points of the four sides.

(8) $r = \frac{2}{3}R$, $r = \dfrac{R}{2}$, no max.

(9) $r = R\sqrt{\dfrac{2}{3}}$, $r = \dfrac{R}{\sqrt{2}}$, $r = 0.8506R$.

(10) At the rate of $\dfrac{8}{r}$ square feet per second.

(11) $r = \dfrac{R\sqrt{8}}{3}$. (12) $n = \sqrt{\dfrac{NR}{r}}$.

Exercises X. (p. 418.)

(1) Max.: $x = -2.19$, $y = 24.19$; min.:, $x = 1.52$, $y = -1.38$.

(2) $\dfrac{dy}{dx} = \dfrac{b}{a} - 2cx$; $\dfrac{d^2y}{dx^2} = -2c$; $x = \dfrac{b}{2ac}$ (*a maximum*).

(3) (*a*) One maximum and two minima.
(*b*) One maximum. ($x = 0$; other points unreal.)

(4) Min.: $x = 1.71$, $y = 6.14$. (5) Max: $x = -.5$, $y = 4$.

(6) Max.: $x = 1.414$, $y = 1.7675$.
Min.: $x = -1.414$, $y = 1.7675$.

(7) Max.: $x = -3.565$, $y = 2.12$.
Min.: $x = +3.565$, $y = 7.88$.

(8) $0.4N$, $0.6N$. (9) $x = \sqrt{\dfrac{a}{c}}$.

(10) Speed 8.66 nautical miles per hour. Time taken 115.47 hours.
Minimum cost \$$112.\ 12$*s*.

(11) Max. and min. for $x = 7.5$, $y = \pm 5.414$.

(12) Min.: $x = \frac{1}{2}$, $y = 0.25$; max.: $x = -\frac{1}{3}$, $y = 1.408$.

Exercises XI. (p. 420.)

(1) $\dfrac{2}{x-3}+\dfrac{1}{x+4}$. (2) $\dfrac{1}{x-1}+\dfrac{2}{x-2}$. (3) $\dfrac{2}{x-3}+\dfrac{1}{x+4}$.

(4) $\dfrac{5}{x-4}-\dfrac{4}{x-3}$. (5) $\dfrac{19}{13(2x+3)}-\dfrac{22}{13(3x-2)}$.

(6) $\dfrac{2}{x-2}+\dfrac{4}{x-3}-\dfrac{5}{x-4}$.

(7) $\dfrac{1}{6(x-1)}+\dfrac{11}{15(x+2)}+\dfrac{1}{10(x-3)}$.

(8) $\dfrac{7}{9(3x+1)}+\dfrac{71}{63(3x-2)}-\dfrac{5}{7(2x+1)}$.

(9) $\dfrac{1}{3(x-1)}+\dfrac{2x+1}{3(x^2+x+1)}$. (10) $x+\dfrac{2}{3(x+1)}+\dfrac{1-2x}{3(x^2-x+1)}$.

(11) $\dfrac{3}{(x+1)}+\dfrac{2x+1}{x^2+x+1}$. (12) $\dfrac{1}{x-1}-\dfrac{1}{x-2}+\dfrac{2}{(x-2)^2}$.

(13) $\dfrac{1}{4(x-1)}-\dfrac{1}{4(x+1)}+\dfrac{1}{2(x+1)^2}$.

(14) $\dfrac{4}{9(x-1)}-\dfrac{4}{9(x+2)}-\dfrac{1}{3(x+2)^2}$.

(15) $\dfrac{1}{x+2}-\dfrac{x-1}{x^2+x+1}-\dfrac{1}{(x^2+x+1)^2}$.

(16) $\dfrac{5}{x+4}-\dfrac{32}{(x+4)^2}+\dfrac{36}{(x+4)^3}$.

(17) $\dfrac{7}{9(3x-2)^2}+\dfrac{55}{9(3x-2)^3}+\dfrac{73}{9(3x-2)^4}$.

(18) $\dfrac{1}{6(x-2)}+\dfrac{1}{3(x-2)^2}-\dfrac{x}{6(x^2+2x+4)}$.

Exercises XII. (p. 421.)

(1) $ab(e^{ax}+e^{-ax})$. (2) $2at+\frac{2}{t}$. (3) $\ln n$.

(5) npv^{n-1}. (6) $\frac{n}{x}$. (7) $\frac{3e^{-\frac{x}{x-1}}}{(x-1)^2}$.

(8) $6xe^{-5x}-5(3x^2+1)e^{-5x}$. (9) $\frac{ax^{a-1}}{x^a+a}$.

(10) $\left(\frac{6x}{3x^2-1}+\frac{1}{2\left(\sqrt{x}+x\right)}\right)\left(3x^2-1\right)\left(\sqrt{x}+1\right)$.

(11) $\frac{1-\ln\left(x+3\right)}{\left(x+3\right)^2}$.

(12) $a^x\left(ax^{a-1}+x^a\ln a\right)$. (14) Min.: $y=0.7$ for $x=0.694$.

(15) $\frac{1+x}{x}$. (16) $\frac{3}{x}(\ln ax)^2$.

Exercises XIII. (p. 422.)

(1) Let $\frac{t}{T}=x$ $(t=8x)$.

(2) $T=34.627$; 159.46 minutes.

(3) Take $2t=x$.

(5) (*a*) $x^x\left(1+\ln x\right)$; (*b*) $2x(e^x)^x$; (*c*) $e^{x^x}\times x^x\left(1+\ln x\right)$.

(6) 0.14 second. (7) (*a*) 1.642; (*b*) 15.58.

(8) $\mu=0.00037$, $31^m\frac{1}{4}$.

(9) i is 63.4% of i_0, 220 kilometres.

(10) 0.133, 0.145, 0.155, mean 0.144; -10.2%, -0.9%, $+77.2\%$.

(11) Min. for $x = \dfrac{1}{e}$.

(12) Max. for $x = e$.

(13) Min. for $x = \ln a$.

Exercises XIV. (p. 423.)

(1) (i) $\dfrac{dy}{d\theta} = A\cos\left(\theta - \dfrac{\pi}{2}\right)$;

(ii) $\dfrac{dy}{d\theta} = 2\sin\theta\cos\theta = \sin 2\theta$ and $\dfrac{dy}{d\theta} = 2\cos 2\theta$;

(iii) $\dfrac{dy}{d\theta} = 3\sin^2\theta\cos\theta$ and $\dfrac{dy}{d\theta} = 3\cos 3\theta$.

(2) $\theta = 45°$ or $\dfrac{\pi}{4}$ radians.

(3) $\dfrac{dy}{dt} = -n\sin 2\pi nt$.

(4) $a^x \ln a \cos a^x$.

(5) $\dfrac{\cos x}{\sin x} = \operatorname{cotan} x$

(6) $18.2\cos(x + 26°)$.

(7) The slope is $\dfrac{dy}{d\theta} = 100\cos(\theta - 15°)$, which is a maximum when $(\theta - 15°) = 0$, or $\theta = 15°$; the value of the slope being then $= 100$. When $\theta = 75°$ the slope is $100\cos(75° - 15°) = 100\cos 60° = 100 \times \frac{1}{2} = 50$.

(8)
$$\begin{aligned}\cos\theta\sin 2\theta + 2\cos 2\theta\sin\theta &= 2\sin\theta\left(\cos^2\theta + \cos 2\theta\right)\\ &= 2\sin\theta\left(3\cos^2\theta - 1\right).\end{aligned}$$

(9) $amn\theta^{n-1}\tan^{m-1}(\theta^n)\sec^2\theta^n$.

(10) $e^x\left(\sin^2 x + \sin 2x\right)$; $e^x\left(\sin^2 x + 2\sin 2x + 2\cos 2x\right)$.

(11) (i) $\dfrac{dy}{dx} = \dfrac{ab}{(x+b)^2}$; (ii) $\dfrac{a}{b}e^{-\frac{x}{b}}$; (iii) $\dfrac{1}{90}^{\circ} \times \dfrac{ab}{(b^2 + x^2)}$.

(12) (i) $\dfrac{dy}{dx} = \sec x \tan x$;

(ii) $\dfrac{dy}{dx} = -\dfrac{1}{\sqrt{1-x^2}}$;

(iii) $\dfrac{dy}{dx} = \dfrac{1}{1+x^2}$;

(iv) $\dfrac{dy}{dx} = \dfrac{1}{x\sqrt{x^2-1}}$;

(v) $\dfrac{dy}{dx} = \dfrac{\sqrt{3\sec x}\left(3\sec^2 x - 1\right)}{2}$.

(13) $\dfrac{dy}{d\theta} = 4.6\,(2\theta+3)^{1.3}\cos(2\theta+3)^{2.3}$.

(14) $\dfrac{dy}{d\theta} = 3\theta^2 + 3\cos(\theta+3) - \ln 3\left(\cos\theta \times 3^{\sin\theta} + 3\theta\right)$.

(15) $\theta = \cot\theta; \theta = \pm 0.86$; is max. for $+\theta$, min. for $-\theta$.

Exercises XVI. (p. 425.)

(1) $1\frac{1}{3}$.　(2) 0.6344.　(3) 0.2624.

(4) (*a*) $y = \frac{1}{8}x^2 + C$; (*b*) $y = \sin x + C$.

(5) $y = x^2 + 3x + C$.

Exercises XVII. (p. 425.)

(1) $\dfrac{4\sqrt{a}x^{\frac{3}{2}}}{3} + C.$ (2) $-\dfrac{1}{x^3} + C.$ (3) $\dfrac{x^4}{4a} + C.$

(4) $\frac{1}{3}x^3 + ax + C.$ (5) $-2x^{-\frac{5}{2}} + C.$

(6) $x^4 + x^3 + x^2 + x + C.$ (7) $\dfrac{ax^2}{4} + \dfrac{bx^3}{9} + \dfrac{cx^4}{16} + C.$

(8) $\dfrac{x^2+a}{x+a} = x - a + \dfrac{a^2+a}{x+a}$ by division. Therefore the answer is $\dfrac{x^2}{2} - ax + (a^2 + a)\ln(x+a) + C.$

(9) $\dfrac{x^4}{4} + 3x^3 + \dfrac{27}{2}x^2 + 27x + C.$ (10) $\dfrac{x^3}{3} + \dfrac{2-a}{2}x^2 - 2ax + C.$

(11) $a^2(2x^{\frac{3}{2}} + \frac{9}{4}x^{\frac{4}{3}}) + C.$ (12) $-\frac{1}{3}\cos\theta - \frac{1}{6}\theta + C.$

(13) $\dfrac{\theta}{2} + \dfrac{\sin 2a\theta}{4a} + C.$ (14) $\dfrac{\theta}{2} - \dfrac{\sin 2\theta}{4} + C.$

(15) $\dfrac{\theta}{2} - \dfrac{\sin 2a\theta}{4a} + C.$ (16) $\frac{1}{3}e^{3x}.$

(17) $\log(1+x) + C.$ (18) $-\ln(1-x) + C.$

Exercises XVIII. (p. 426.)

(1) Area $= 60$; mean ordinate $= 10$.

(2) Area $= \frac{2}{3}$ of $a \times 2a\sqrt{a}$.

(3) Area $= 2$; mean ordinate $= \dfrac{2}{\pi} = 0.637$.

(4) Area $= 1.57$; mean ordinate $= 0.5$.

(5) 0.572, 0.0476.

(6) Volume $= \pi r^2 \dfrac{h}{3}$.

(7) 1.25.

(8) 79.4.

(9) Volume $= 4.9348$; area of surface $= 12.57$ (from 0 to π).

(10) $a \ln a$, $\quad \dfrac{a}{a-1} \ln a$.

(12) Arithmetical mean $= 9.5$; quadratic mean $= 10.85$.

(13) Quadratic mean $= \dfrac{1}{\sqrt{2}} \sqrt{A_1^2 + A_3^2}$; arithmetical mean $= 0$.

The first involves a somewhat difficult integral, and may be stated thus: By definition the quadratic mean will be

$$\sqrt{\frac{1}{2\pi} \int_0^{2\pi} (A_1 \sin x + A_3 \sin 3x)^2 \, dx.}$$

Now the integration indicated by

$$\int (A_1^2 \sin^2 x + 2A_1 A_3 \sin x \sin 3x + A_3^2 \sin^2 3x) \, dx$$

is more readily obtained if for $\sin^2 x$ we write

$$\frac{1 - \cos 2x}{2}.$$

For $2 \sin x \sin 3x$ we write $\cos 2x - \cos 4x$; and, for $\sin^2 3x$,

$$\frac{1 - \cos 6x}{2}.$$

Making these substitutions, and integrating, we get

$$\frac{A_1^2}{2} \left(x - \frac{\sin 2x}{2} \right) + A_1 A_3 \left(\frac{\sin 2x}{2} - \frac{\sin 4x}{4} \right) + \frac{A_3^2}{2} \left(x - \frac{\sin 6x}{6} \right).$$

At the lower limit the substitution of 0 for x causes all this to vanish, whilst at the upper limit the substitution of 2π for x gives $A_1^2 \pi + A_3^2 \pi$. And hence the answer follows.

(14) Area is 62.6 square units. Mean ordinate is 10.42.

(16) 436.3. (This solid is pear shaped.)

Exercises XIX. (p. 427.)

(1) $\dfrac{x\sqrt{a^2-x^2}}{2}+\dfrac{a^2}{2}\sin^{-1}\dfrac{x}{a}+C.$

(2) $\dfrac{x^2}{2}(\ln x-\frac{1}{2})+C.$

(3) $\dfrac{x^{a+1}}{a+1}\left(\ln x-\dfrac{1}{a+1}\right)+C.$

(4) $\sin e^x+C.$

(5) $\sin(\ln x)+C.$

(6) $e^x(x^2-2x+2)+C.$

(7) $\dfrac{1}{a+1}(\ln x)^{a+1}+C.$

(8) $\ln(\ln x)+C.$

(9) $2\ln(x-1)+3\ln(x+2)+C.$

(10) $\frac{1}{2}\ln(x-1)+\frac{1}{5}\ln(x-2)+\frac{3}{10}\ln(x+3)+C.$

(11) $\dfrac{b}{2a}\ln\dfrac{x-a}{x+a}+C.$

(12) $\ln\dfrac{x^2-1}{x^2+1}+C.$

(13) $\frac{1}{4}\ln\dfrac{1+x}{1-x}+\frac{1}{2}\tan^{-1}(x)+C.$

(14) $\dfrac{1}{\sqrt{a}}\log\dfrac{\sqrt{a}-\sqrt{a-bx^2}}{x\sqrt{a}}$. (Let $\dfrac{1}{x}=v$; then, in the result, let $\sqrt{v^2-\dfrac{b}{a}}=v-u$.)

You had better differentiate now the answer and work back to the given expression as a check.

Appendix A

Constants, units, and conversion ratios

It this appendix you will find a number of tables of useful information which you might need when solving math and physics problems.

Fundamental constants of Nature

Many of the equations of physics include constants as parameters of the equation. For example, Newton's law of gravitation says that the force of gravity between two objects of mass M and m separated by a distance r is $F_g = \frac{GMm}{r^2}$, where G is Newton's gravitational constant.

Symbol	Value	Units	Name
G	$6.673\,84 \times 10^{-11}$	$\mathrm{m^3\,kg^{-1}s^{-2}}$	gravitational constant
g	$9.806\,65 \approx 9.81$	$\mathrm{m\,s^{-2}}$	Earth free-fall acceleration
m_p	$1.672\,621 \times 10^{-27}$	kg	proton mass
m_e	$9.109\,382 \times 10^{-31}$	kg	electron mass
N_A	$6.022\,141 \times 10^{23}$	$\mathrm{mol^{-1}}$	Avogadro's number
k_B	$1.380\,648 \times 10^{-23}$	$\mathrm{J\,K^{-1}}$	Boltzmann's constant
R	$8.314\,462\,1$	$\mathrm{J\,K^{-1}\,mol^{-1}}$	gas constant $R = N_\mathrm{A}k_\mathrm{B}$
μ_0	$1.256\,637 \times 10^{-6}$	$\mathrm{N\,A^{-2}}$	permeability of free space
ε_0	$8.854\,187 \times 10^{-12}$	$\mathrm{F\,m^{-1}}$	permittivity of free space
c	$299\,792\,458$	$\mathrm{m\,s^{-1}}$	speed of light $c = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$
e	$1.602\,176 \times 10^{-19}$	C	elementary charge
h	$6.626\,069 \times 10^{-34}$	$\mathrm{J\,s}$	Planck's constant

Units

The International System of Units (*Système International*) defines seven base units for measuring physical quantities.

Name	Sym.	Measures	Definition
metre	m	length	The distance travelled by light in vacuum during $\frac{1}{299792458^{\text{th}}}$ of a second.
kilogram	kg	mass	The mass of the *international prototype kilogram* (a cylinder of platinum-iridium kept at Sèvres near Paris).
second	s	time	The time for 9192631770 transitions in the ground state of the caesium-133 atom.
Ampere	A	electric current	One ampere is the current that has to flow in two infinitely long wires placed a distance 1[m] apart, to produce a force between them of 2×10^{-7}[N/m].
Kelvin	K	temperature	The Kelvin is $\frac{1}{273.16}$ of the thermodynamic temperature of the triple point of water.
mole	mol	# of atoms	One mole is how many carbon atoms are in 0.012[kg] of carbon-12.
candela	cd	light intensity	One candela is defined as the luminous intensity of a monochromatic source with a particular frequency and radiant intensity.

Derived units

The base SI units cover most of the fundamental quantities. Other physical units are defined as combinations of the basic units.

Name	Sym.	Measures	Definition	SI equivalent
Hertz	Hz	frequency		s^{-1}
Newton	N	force		$\mathrm{kg\,m\,s}^{-2}$
Pascal	Pa	pressure	$\mathrm{N/m}^2$	$\mathrm{kg\,m}^{-1}\,\mathrm{s}^{-2}$
Joule	J	energy, work, heat	N m	$\mathrm{kg\,m}^2\,\mathrm{s}^{-2}$
Watt	W	power	J/s	$\mathrm{kg\,m}^2\,\mathrm{s}^{-3}$
Coulomb	C	electric charge		s A
Volt	V	voltage, electric potential		$\mathrm{kg\,m}^2\,\mathrm{s}^{-3}\,\mathrm{A}^{-1}$
Ohm	Ω	resistance, reactance	V/A	$\mathrm{kg\,m}^2\,\mathrm{s}^{-3}\,\mathrm{A}^{-2}$
Siemens	S	electrical conductance	A/V	$\mathrm{kg}^{-1}\,\mathrm{m}^{-2}\,\mathrm{s}^3\,\mathrm{A}^2$
Farad	F	capacitance	C/V	$\mathrm{kg}^{-1}\,\mathrm{m}^{-2}\,\mathrm{s}^4\,\mathrm{A}^2$
Tesla	T	magnetic field strength		$\mathrm{kg\,s}^{-2}\,\mathrm{A}^{-1}$
Henry	H	inductance	Ω s	$\mathrm{kg\,m}^2\,\mathrm{s}^{-2}\,\mathrm{A}^{-2}$
Weber	Wb	magnetic flux	$\mathrm{T\,m}^2$	$\mathrm{kg\,m}^2\,\mathrm{s}^{-2}\,\mathrm{A}^{-1}$

Other units and conversions

We often measure physical quantities like length, weight, and velocity in nonstandard units like feet, pounds, and miles per hour. The following table lists the conversion ratios which are required to covert these nonstandard measurement units to SI units.

Dimension	Symb.	Name	Conversion
length	Å	Angstrom	$1[\text{Å}] = 10^{-10}[\mathrm{m}] = 0.1[\mathrm{nm}]$
	in, "	inch	$1[\mathrm{T}] = 1000[\mathrm{kg}]$
	ft, '	foot	$1[\mathrm{ft}] = 12[\mathrm{in}] = 0.3048[\mathrm{m}]$
	yd	yard	$1[\mathrm{yd}]= 3[\mathrm{ft}] = 0.9144[\mathrm{m}]$
	mi	mile	$1[\mathrm{mi}] = 5280[\mathrm{ft}] = 1609.344[\mathrm{m}]$
	nmi	nautical mile	$1[\mathrm{nmi}] = 1852[\mathrm{m}]$
	ly	light-year	$1[\mathrm{ly}] = 9.460\,730\,472 \times 10^{15}[\mathrm{m}]$
area	in^2	square inch	$1[\mathrm{in}^2] = 6.452 \times 10^{-4}[\mathrm{m}^2]$
	ft^2	square foot	$1[\mathrm{ft}^2] = 9.290 \times 10^{-2}[\mathrm{m}^2]$
			$1[\mathrm{m}^2] = 110.764[\mathrm{ft}^2]$
	ac	acre	$1[\mathrm{ac}] = 4840[\mathrm{yd}^2] = 4046.856[\mathrm{m}^2]$
	ha	hectare	$1[\mathrm{ha}] = 10\,000[\mathrm{m}^2]$
	mi^2	square mile	$1[\mathrm{mi}^2] = 2.589\,988 \times 10^{6}[\mathrm{m}^2]$
volume	L	litre	$1[\mathrm{L}] = 1[\mathrm{dm}^3] = \frac{1}{1000}[\mathrm{m}^3]$
	gal(US)	gallon (fluid)	$1[\mathrm{gal}] = 3.785[\mathrm{L}]$
weight	lb	pound	$1[\mathrm{lb}] = 0.454[\mathrm{kg}] = 453.592[\mathrm{g}]$
	t	tonne	$1[\mathrm{t}] = 1000[\mathrm{kg}]$
angle	rad	radian	$1[\mathrm{turn}] = 2\pi[\mathrm{rad}]$
	°	degree	$360[^\circ] = 2\pi[\mathrm{rad}]$
	rev	revolution	$1[\mathrm{rev}] = 360[^\circ] = 2\pi[\mathrm{rad}]$
	grad	gradian	$1[\mathrm{grad}]=\frac{1}{400}[\mathrm{rev}] = 0.9[^\circ]$
time	min	minute	$1[\mathrm{min}] = 60[\mathrm{s}]$
	h	hour	$1[\mathrm{h}] = 60[\mathrm{min}] = 3600[\mathrm{s}]$
velocity	km/h	km per hour	$1[\mathrm{km/h}] = \frac{1}{3.6}[\mathrm{m/s}] = 0.2\overline{7}[\mathrm{m/s}]$
	mph	mile per hour	$1[\mathrm{mph}] = 0.447[\mathrm{m/s}] = 1.61[\mathrm{km/h}]$
temperature	°C	Celsius	$x[^\circ\mathrm{C}] = (x + 273.15)[^\circ\mathrm{K}]$
	°F	Fahrenheit	$x[^\circ\mathrm{F}] = \frac{5}{9}(x + 459.67)[^\circ\mathrm{K}]$
			$x[^\circ\mathrm{F}] = \frac{5}{9}(x - 32)[^\circ\mathrm{C}]$
pressure	atm	atmosphere	$1[\mathrm{atm}] = 101\,325[\mathrm{Pa}]$
	bar	bar	$1[\mathrm{bar}] =10^{5}[\mathrm{Pa}]$

Appendix B

Notation

This appendix contains a summary of the notation used in this book.

Math notation

Expression	Read as	Used to
a, b, x, y		denote variables
$=$	is equal to	indicate two expressions are equal in value
$\equiv$	is defined as	define a variable in terms of an expression
$a + b$	a plus b	combine lengths
$a - b$	a minus b	find the difference in length
$a \times b \equiv ab$	a times b	find the area of a rectangle
$a^2 \equiv aa$	a squared	find the area of a square of side length a
$a^3 \equiv aaa$	a cubed	find the volume of a cube of side length a
a^n	a exponent n	denote a multiplied by itself n times
$\sqrt{a} \equiv a^{\frac{1}{2}}$	square root of a	find the side length of a square of area a
$\sqrt[3]{a} \equiv a^{\frac{1}{3}}$	cube root of a	find the side of a cube with volume a
$a/b \equiv \frac{a}{b}$	a divided by b	denote parts of a whole
$a^{-1} \equiv \frac{1}{a}$	one over a	denotes division by a
$f(x)$	f of x	denote the output of the function f applied to the input x
f^{-1}	f inverse	denote the inverse function of $f(x)$ if $f(x) = y$, then $f^{-1}(y) = x$
e^x	e to the x	denote the exponential function base e
$\ln(x)$	natural log of x	logarithm base e
a^x	a to the x	denote the exponential function base a
$\log_a(x)$	log base a of x	logarithm base a
θ, ϕ	*theta*, *phi*	denote angles
$\sin, \cos, \tan$	sin, cos, tan	obtain trigonometric ratios
$\%$	percent	denote proportions of a total $a\% \equiv \frac{a}{100}$

Set notation

You don't need a lot of fancy notation to understand mathematics. It really helps, though, if you know a little bit of set notation.

Symbol	Read as	Denotes
$\{ \dots \}$	the set ...	define a sets
$\vert$	such that	describe or restrict the elements of a set
$\mathbb{N}$	the naturals	the set $\mathbb{N} \equiv \{0, 1, 2, 3, \ldots\}$
$\mathbb{Z}$	the integers	the set $\mathbb{Z} \equiv \{\ldots, -2, -1, 0, 1, 2, 3, \ldots\}$
$\mathbb{Q}$	the rationals	the set of fractions of integers
$\mathbb{A}$		the set of algebraic numbers
$\mathbb{R}$		the set of real numbers
$\mathbb{C}$		the set of complex numbers
$\subset$	subset	one set contained in another
$\subseteq$	subset or equal	containment or equality
$\cup$	union	the combined elements from two sets
$\cap$	intersection	the elements two sets have in common
$S \setminus T$	S set minus T	the elements of S that are not in T
$a \in S$	a in S	a is an element of set S
$a \notin S$	a not in S	a is not an element of set S
$\forall x$	for all x	a statement that holds for all x
$\exists x$	there exists x	an existence statement
$\nexists x$	there doesn't exist x	a non-existence statement

Vectors notation

Expression	Denotes
$\vec{v}$	a vector
(v_x, v_y)	vector in component notation
$v_x\hat{\imath} + v_y\hat{\jmath}$	vector in unit vector notation
$\|\vec{v}\|\angle\theta$	vector in length-and-direction notation
$\|\vec{v}\|$	length of the vector $\vec{v}$
θ	angle the vector $\vec{v}$ makes with the x-axis
$\hat{v} \equiv \frac{\vec{v}}{\|\vec{v}\|}$	unit length vector in the same direction as $\vec{v}$
$\vec{u} \cdot \vec{v}$	dot product of the vectors $\vec{u}$ and $\vec{v}$
$\vec{u} \times \vec{v}$	cross product of the vectors $\vec{u}$ and $\vec{v}$

Mechanics notation

Expression	Denotes
$\vec{F}$	a force
m	mass of an object
$a(t)$	acceleration of an object as a function of time
$v(t)$	velocity of an object as a function of time
$x(t)$	position of an object as a function of time
$\vec{N}$	normal force
$\vec{F}_{fs}$	static force of friction
$\vec{F}_{fk}$	kinetic force of friction
$\vec{F}_g \equiv \vec{W}$	gravitational force; the weight of an object
U_g	gravitational potential energy
$\vec{F}_s$	force of a spring
U_s	spring potential energy
$\vec{p}$	momentum of a moving object
K	kinetic energy
$\mathcal{T}$	torque
I_{obj}	moment of inertia of an object
$\alpha(t)$	angular acceleration of an object as a function of time
$\omega(t)$	angular velocity of an object as a function of time
$\theta(t)$	angular position of an object as a function of time
L	angular momentum of a spinning object
K_r	rotational kinetic energy of a spinning object

Calculus notation

Expression	Denotes
∞	infinity
ϵ, δ	the Greek letters *epsilon* and *delta*
$f(x)$	a function
$\lim_{x\to\infty} f(x)$	limit of $f(x)$ as x goes to infinity
$\lim_{x\to a} f(x)$	limit of $f(x)$ as x goes to a
$f'(x)$	derivative of $f(x)$
$f''(x)$	second derivative of $f(x)$
$\frac{d}{dx}$	derivative operator
$F(x)$	antiderivative function of $f(x)$
$\int f(x)\,dx$	indefinite integral of $f(x)$
$\int_a^b f(x)\,dx$	definite integral of $f(x)$ between $x=a$ and $x=b$
$F(x)\big\rvert_\alpha^\beta$	change in $F(x)$ between α and β: $F(x)\big\rvert_\alpha^\beta = F(\beta) - F(\alpha)$
a_n	a sequence $a_n : \mathbb{N} \to \mathbb{R}$
$\sum_{n=0}^\infty a_n$	infinite series of the sequence a_n

Appendix C

Formulas

Calculus formulas

$\dfrac{dy}{dx}$	$\longleftarrow \quad y \quad \longrightarrow$	$\displaystyle\int y\,dx$
Algebraic		
1	x	$\frac{1}{2}x^2 + C$
0	a	$ax + C$
nx^{n-1}	x^n	$\dfrac{1}{n+1}x^{n+1} + C$
$-x^{-2}$	x^{-1}	$\ln x + C$
$\dfrac{du}{dx} \pm \dfrac{dv}{dx} \pm \dfrac{dw}{dx}$	$u \pm v \pm w$	$\int u\,dx \pm \int v\,dx \pm \int w\,dx$
$u\,\dfrac{dv}{dx} + v\,\dfrac{du}{dx}$	uv	No general form known
$\dfrac{v\,\dfrac{du}{dx} - u\,\dfrac{dv}{dx}}{v^2}$	$\dfrac{u}{v}$	No general form known
$\dfrac{du}{dx}$	u	$ux - \int x\,du + C$

$\dfrac{dy}{dx}$	$\longleftarrow \quad y \quad \longrightarrow$	$\displaystyle\int y\,dx$
Exponential and Logarithmic		
e^x	e^x	$e^x + C$
x^{-1}	$\ln x$	$x(\ln x - 1) + C$
$0.4343 \times x^{-1}$	$\log_{10} x$	$0.4343x(\ln x - 1) + C$
$a^x \ln a$	a^x	$\dfrac{a^x}{\ln a} + C$
Trigonometric		
$\cos x$	$\sin x$	$-\cos x + C$
$-\sin x$	$\cos x$	$\sin x + C$
$\sec^2 x$	$\tan x$	$-\ln \cos x + C$
Circular (Inverse)		
$\dfrac{1}{\sqrt{(1-x^2)}}$	$\sin^{-1}(x)$	$x \sin^{-1}(x) + \sqrt{1-x^2} + C$
$-\dfrac{1}{\sqrt{(1-x^2)}}$	$\cos^{-1}(x)$	$x \cos^{-1}(x) - \sqrt{1-x^2} + C$
$\dfrac{1}{1+x^2}$	$\tan^{-1}(x)$	$x \tan^{-1}(x) - \frac{1}{2} \ln(1+x^2) + C$
Hyperbolic		
$\cosh x$	$\sinh x$	$\cosh x + C$
$\sinh x$	$\cosh x$	$\sinh x + C$
$\operatorname{sech}^2 x$	$\tanh x$	$\ln \cosh x + C$

$\dfrac{dy}{dx}$	$\longleftarrow \quad y \quad \longrightarrow$	$\displaystyle\int y\,dx$
Miscellaneous		
$-\dfrac{1}{(x+a)^2}$	$\dfrac{1}{x+a}$	$\ln(x+a)+C$
$-\dfrac{x}{(a^2+x^2)^{\frac{3}{2}}}$	$\dfrac{1}{\sqrt{a^2+x^2}}$	$\ln(x+\sqrt{a^2+x^2})+C$
$\mp\dfrac{b}{(a\pm bx)^2}$	$\dfrac{1}{a\pm bx}$	$\pm\dfrac{1}{b}\ln(a\pm bx)+C$
$-\dfrac{3a^2x}{(a^2+x^2)^{\frac{5}{2}}}$	$\dfrac{a^2}{(a^2+x^2)^{\frac{3}{2}}}$	$\dfrac{x}{\sqrt{a^2+x^2}}+C$
$a\cos ax$	$\sin ax$	$-\dfrac{1}{a}\cos ax+C$
$-a\sin ax$	$\cos ax$	$\dfrac{1}{a}\sin ax+C$
$a\sec^2 ax$	$\tan ax$	$-\dfrac{1}{a}\ln\cos ax+C$
$\sin 2x$	$\sin^2 x$	$\dfrac{x}{2}-\dfrac{\sin 2x}{4}+C$
$-\sin 2x$	$\cos^2 x$	$\dfrac{x}{2}+\dfrac{\sin 2x}{4}+C$
$n\sin^{n-1}x\cos x$	$\sin^n x$	$-\dfrac{\cos x}{n}\sin^{n-1}x+\dfrac{n-1}{n}\displaystyle\int\sin^{n-2}x\,dx+C$
$-\dfrac{\cos x}{\sin^2 x}$	$\dfrac{1}{\sin x}$	$\ln\tan\dfrac{x}{2}+C$
$-\dfrac{\sin 2x}{\sin^4 x}$	$\dfrac{1}{\sin^2 x}$	$-\operatorname{cotan}x+C$
$\dfrac{\sin^2 x-\cos^2 x}{\sin^2 x\cos^2 x}$	$\dfrac{1}{\sin x\cos x}$	$\ln\tan x+C$
$n\sin mx\cos nx+$ $m\sin nx\cos mx$	$\sin mx\sin nx$	$\frac{1}{2}\cos(m-n)x-\frac{1}{2}\cos(m+n)x+C$
$2a\sin 2ax$	$\sin^2 ax$	$\dfrac{x}{2}-\dfrac{\sin 2ax}{4a}+C$
$-2a\sin 2ax$	$\cos^2 ax$	$\dfrac{x}{2}+\dfrac{\sin 2ax}{4a}+C$

Mechanics formulas

Forces:

$$W = F_g = \frac{GMm}{r^2} = gm, \quad F_s = -kx, \quad F_{fs} \leq \mu_s N, \quad F_{fk} = \mu_k N$$

Newton's laws:

$$\text{if no } \vec{F}_{\text{ext}}, \text{ then } \vec{v}_i = \vec{v}_f \tag{1}$$

$$\vec{F}_{\text{net}} = m\vec{a} \tag{2}$$

$$\text{if } \vec{F}_{12}, \text{ then } \exists \vec{F}_{21} = -\vec{F}_{12} \tag{3}$$

Uniform acceleration motion (UAM):

$$a(t) = a \tag{4}$$

$$v(t) = at + v_i \tag{5}$$

$$x(t) = \tfrac{1}{2}at^2 + v_i t + x_i \tag{6}$$

$$v_f^2 = v_i^2 + 2a\Delta x \tag{7}$$

Momentum:

$$\vec{p} = m\vec{v} \tag{8}$$

Energy and work:

$$K = \tfrac{1}{2}mv^2, \quad U_g = mgh, \quad U_s = \tfrac{1}{2}kx^2, \quad K_r = \tfrac{1}{2}I\omega^2, \quad W = \vec{F} \cdot \vec{d} \tag{9}$$

Conservation laws:

$$\sum \vec{p}_{\text{in}} = \sum \vec{p}_{\text{out}} \tag{10}$$

$$L_{\text{in}} = L_{\text{out}} \tag{11}$$

$$\sum E_{\text{in}} + W_{\text{in}} = \sum E_{\text{out}} + W_{\text{out}} \tag{12}$$

Circular motion (radial acceleration and radial force):

$$a_r = \frac{v_t^2}{R}, \qquad F_r = ma_r \tag{13}$$

Angular motion:

$$\begin{aligned} F = ma &\quad\Rightarrow\quad \mathcal{T} = I\alpha && (14)\\ a(t), v(t), x(t) &\quad\Rightarrow\quad \alpha(t), \omega(t), \theta(t) && (15)\\ \vec{p} = m\vec{v} &\quad\Rightarrow\quad L = I\omega && (16)\\ K = \tfrac{1}{2}mv^2 &\quad\Rightarrow\quad K_r = \tfrac{1}{2}I\omega^2 && (17) \end{aligned}$$

SHM with $\omega = \sqrt{\frac{k}{m}}$ (mass-spring system) or $\omega = \sqrt{\frac{g}{\ell}}$ (pendulum):

$$\begin{aligned} x(t) &= A\cos(\omega t + \phi) && (18)\\ v(t) &= -A\omega\sin(\omega t + \phi) && (19)\\ a(t) &= -A\omega^2\cos(\omega t + \phi) && (20) \end{aligned}$$